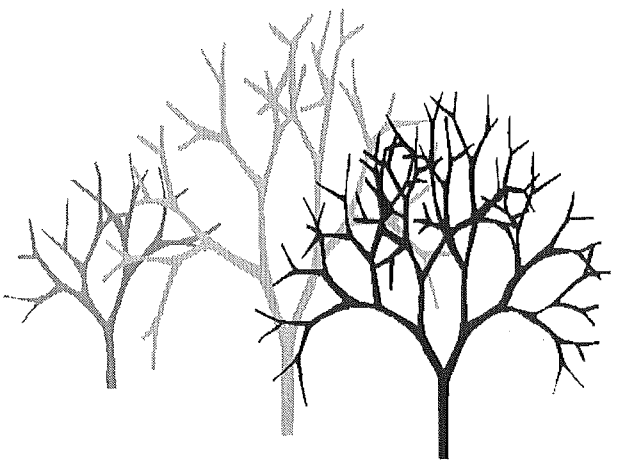
University of Canterbury
Department of Mathematics and Statistics

Stochastic Speciation Models For Evolutionary Trees

A thesis submitted in partial fulfilment
of the requirements for
the Degree of
Doctor of Philosophy in Mathematics
at the
University of Canterbury
by

Andy McKenzie

Supervisor: Assoc. Prof. Mike Steel
2000



Abstract

Phylogenetic trees are widely used in biology to represent evolutionary relationships between species. As the details of the evolutionary process are mostly unknown, modelling work on the shapes of these trees has had to incorporate a random component. Two null models introduced for this purpose are the uniform model and the Yule model. A third model, the comb model, is useful for giving bounds on theoretical results. We investigate some mathematical properties of these three models.

Let the *distance* between two nodes be the number of edges separating them. We find exact formulae for the mean distance of a randomly chosen leaf from the root, and for the mean distance between two randomly chosen leaves of a rooted tree. In addition, for the Yule model we find the probability distribution for the distance of randomly chosen leaf from the root.

A *cherry* is a pair of leaves which are adjacent to a common node. By realising the process of cherry formation by extended Polya urn models we show that the number of cherries is asymptotically normal. This allows us to develop simple statistical tests for the Yule and uniform null hypotheses for the growth of rooted trees. A *triplet* is a cherry and a pendant edge that are adjacent to a common node. We also show that the asymptotic distribution of triplets is normal for the Yule model, and put forward a conjecture for the distribution under the uniform model.

The construction of an evolutionary tree is generally a two stage process: an unrooted tree is constructed, then it is rooted. We investigate a method for rooting a tree based on the shape of the tree and the Yule model for the growth of rooted trees. We show that even for trees with large number of leaves the approximate location of the root can be located with high probability.

Let S be a set of two rooted binary trees for which the leaf sets $\mathcal{L}_1, \mathcal{L}_2$ form a partition

of the set $\{1, 2, \ldots, n\}$. We derive a recursion for the number of trees on n leaves that are compatible with the the set S. We extend this recursion for a set S of three trees, but show that the numbers of terms required in the recursion grows at least exponentially with the number of trees in the set S.

Let S be a set of rooted binary trees. A tree which is a subtree of each of the trees in the set is called an agreement subtree, and such a tree with the maximum number of possible leaves is called a maximum agreement subtree (MAST). We derive an upper bound for the probability that two randomly generated trees have a MAST with number of leaves greater than or equal to a given value s. We find the form the upper bound takes when the trees are generated according to the uniform and Yule models.

The *entropy* of a probability distribution is equal to the mean information, where the information of an event E is $-\log \mathbb{P}(E)$. We derive exact and asymptotic formulae for the entropy of the comb, uniform and Yule probability distributions.

We show that the comb, uniform, and Yule models satisfy a property called *group elimination*. A special case of the property of group elimination is *sampling consistency*. We show that for any probability distribution on trees that satisfies sampling consistency there is an upper bound on the probability of the fully symmetric tree shapes.

We introduce a modification of the Yule model in which the speciation rate is a function of the time since the last speciation event of a lineage. Using analytical methods we investigate the probability (conditional and unconditional) of the symmetric tree on four leaves under this modified model. If the speciation rate is constant then the probability of the symmetric tree is the same as in the Yule model. Making the speciation rate zero for a period after a speciation event, then constant afterwards, is found to make the symmetric tree more probable. If the speciation rate is constant for some period after a speciation event, then subsequently zero, the symmetric tree is found to be less probable.

Most of Chapter 3 appeared in the paper "Distribution of cherries for two models of trees" written jointly by Mike Steel and myself [46]. This paper was published in *Mathematical Biosciences*, volume 164, number 1, 2000 (pages 81–92).

Chapter 4 and part of Chapter 2 appeared in the the paper "Properties of phylogenetic trees generated by Yule-type speciation models" written jointly by Mike Steel and myself [68]. This paper is in press in *Mathematical Biosciences*, volume 170, number 1, 2001 (pages 91–112).

The work in the remaining chapters was done under the supervision and guidance of Mike Steel and, unless otherwise indicated, is my own work. Thank you to one of my examiners, Vincent Moulton, for some suggested ammendments which helped to tidy up the thesis.

Andy McKenzie, 20th December, 2000.

Acknowledgements

A big thanks goes out to Mike Steel, my supervisor, for the inspirational and motivational example he has set. I have learnt a lot from Mike and have felt lucky to have had him as a supervisor. Also my gratitude to the good ship Marsden for supplying me with the money to make the journey possible.

For entertaining me, and allowing me to entertain them, I thank all the postgraduate students in the mathematics department. A special thanks to my office mate Paul Shorten for enduring, with a laugh and a smile, my rants on the latest bits of trivia I had found on the web. Also a big thanks to Jon Cherrie, from the office next door, for instigating all manner of frivolity. Cheers also to Chris Hann and Cameron Mouat, part of the five musketeers on the fifth floor.

I particularly appreciated the high drama of indoor soccer each week, where 'Ralph' and 'The Headless Chooks' really kicked some butts, with some occasional lapses. The high point of this was where, in the form of 'Hitmen in Drag' (Paul, Jon, Cameron, and myself), we won the end of term indoor soccer tournament and some prize money!

My other form of physical entertainment was the art and sport of judo. Knowing that someone is trying to throw you through the air, pin you to the ground, or break your arm, helps one to mentally disengage from the mathematical troubles of the day! Thanks to all the judoka at Can-Am-Ju for providing this distraction, and for the recovery sessions at the pub afterwards. A big thanks to my coaches Steve Cooper and Ino Kelderman, and thanks to my grading partner Yolanda Kelderman for all those breakfalls over the years.

Lastly, a bouquet of flowers to the Aorangi Road Brewery and Barbecue Boutique, a haven for the desperate from sobriety and responsibility. Many a fine Friday and Saturday evening was spent there in the company of Chris Gunn, Michael Warman, Grant Rule, Yuka Yanagisawa, and assorted hangerons.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

Evolutionary trees represent the history of speciation for a group of species in a simple visual form. What is immediately apparent from an inspection of these trees is the wide variety of shapes they can take. While the shape of a evolutionary tree is determined by how the processes of speciation and extinction occur, neither speciation or extinction are well understood and are dependent on historical events we may never be able to ascertain. This has encouraged the development of *stochastic* models for evolutionary trees, for example the uniform model and Yule model.

A pertinent debate in this context concerns the relative importance of adaptive and stochastic factors in the process of lineage diversification. Adaptive factors are characteristics of a taxon thought to be responsible for the particular pattern of species survival or extinction observed in that taxon. Stochastic factors are those factors, random in appearance, that independently and uniformly effect the formation and extinction of all species in a tree. If stochastic factors dominate then it can be expected that an appropriate stochastic model should do reasonably well at reproducing the patterns of tree shape observed in evolutionary trees. Conversely, to the extent that a stochastic model is inadequate, this implies that adaptive factors are important, and an inquiry is warranted into the particular form that they take.

Mathematics, particularly probability theory, is the dominant form of analysis used to compare evolutionary trees to those produced by stochastic models. Using mathematics, statistical tests can be developed which quantitatively measure the match between real

trees and those produced in the models. However, this is only part of the role of mathematics. Starting with an appropriate stochastic model, one can investigate the effect on tree shape of changing the model assumptions, explore the viability of new techniques in phylogenetic tree reconstruction, and make links and gain insights from other fields that use similar mathematical models. In short, mathematics offers a precise and systematic language in which to investigate the phenomena at hand.

1.2 Preview

In the following chapters we investigate some mathematical aspects of two simple stochastic models for the growth of rooted trees: uniform and Yule model. We also investigate a third model, the comb model, which is useful for giving bounds on theoretical results. The underlying motivation is to use these models to aid in the understanding of problems in phylogenetic tree reconstruction, though on occasions we take some mathematical detours that are not of any obvious immediate biological significance.

In Chapter 2 we analyse various distance relationships for the Yule and uniform models, where by 'distance' we mean the number of edges separating two nodes. The type of trees we analyse are *rooted trees*, where a rooted tree is one with a root node from which all the other nodes descend as ancestors. In particular, we are concerned with the distance of leaf vertices from the root node, and the distance between two leaves of a rooted tree. These quantities, while having an obvious mathematical appeal, are also useful for estimating distances in real trees in which the edge structure cannot be completely resolved by the biological data (i.e. trees with polytomies).

In both the Yule and uniform models (and many others) every labelled tree on n leaves can be generated, but the probability of the trees differs between models. Because the probability of the trees differ, then so too will the probability distribution of certain characteristics of the trees, and thus the probability distribution of these characteristics can be used to determine the compatibility of a given tree with one of the models. Here we focus on an easily determined characteristic of trees: the number of cherries, where a cherry is a pair of leaves adjacent to a common vertex. In Chapter 3 we show that the distribution of cherries under the Yule and uniform models is asymptotically normal. Using this result we develop statistical tests for the Yule and uniform model hull hypotheses based on the

number of cherries that a tree has. As an extension we also show that the distribution of triplets is asymptotically normal, where a 'triplet' is a cherry and a pendant edge that are adjacent to a common vertex.

Commonly in phylogenetic tree reconstruction one first constructs an unrooted tree, then roots this tree along some edge using some external data. Sometimes this second step, rooting the tree, can be problematic due to the lack of appropriate external data and some heuristic method based only on the tree shape has to be used instead. In Chapter 4 we investigate a maximum likelihood method for rooting a tree based on the shape of the unrooted tree and a simple stochastic model for the growth of rooted trees (the Yule model).

A second issue that may need to be dealt with after the reconstruction of a collection of trees is fitting them together into one single tree (sometimes called a supertree). In Chapter 5 we address the issue of how many supertrees with leaf set $\mathcal{L}$ there are for a set of trees for which the leaf sets form a partition of $\mathcal{L}$. We introduce recursive algorithms for finding this number when there are two trees or three trees. However, we show that the approach used in these algorithms leads to a very large number of terms in the recursions when applied to more than three trees.

A third issue that is relevant after the reconstruction of a set of trees is just what information do they have in common? A tree which is a subtree of each of the trees in the set is called an agreement subtree, and such a tree with the maximum possible number of leaves is called a maximum agreement subtree (MAST). In Chapter 6 we derive an upper bound for the probability that two randomly generated binary trees have a MAST with number of leaves greater than or equal to given value of s. We find the algebraic form the upper bound takes under the Yule and uniform models.

Keeping with the theme of information we look at entropy, a concept with its origin in physics, but which has proven to be useful in communication theory, statistical inference, and complexity theory. Apart from some applications to the study of the DNA code, and some speculative work in evolutionary theory, entropy has yet to find a true home in biology. We ameliorate this situation a little in Chapter 7 by calculating the entropy for the comb, uniform, and Yule models.

In Chapter 8 we look at a property of probability distributions on rooted binary tree called *group elimination*, of which a special case is the *sampling consistency* property.

Group elimination is a property that is of special interest in the context of stochastic models of speciation as the Yule, uniform, and comb models all satisfy group elimination. Furthermore it has been conjectured that they are the only probability distributions on rooted binary trees that do so. We prove that they do indeed satisfy group elimination, and derive an upper bound for the probability of the fully symmetric tree shapes under a probability distribution that satisfies sampling consistency. This work is more concerned with the theoretical background to the modelling process than with any actual biological applications.

Lastly, in Chapter 9, we look at a modification of the Yule model in which the rate of speciation is not the same across all lineages (one of the models basic assumptions). The motivation behind this modification is to ascertain what effect it has on tree balance, knowing that the tree balance of actual evolutionary trees is less than that of the Yule model.

In the rest of this chapter we introduce some basic terminology and formulae relating to trees (Section 1.3). Then we define the Catalan numbers which will be found to be useful in the context of the uniform model (Section 1.4). Following this we define and explain the comb, uniform, and Yule models (Section 1.5). Lastly we look at the empirical match between actual evolutionary trees and the uniform and Yule model trees (Section 1.6).

Throughout we use $f(n) \sim g(n)$ to mean $\lim_{n\to\infty} f(n)/g(n) = 1$. The probability of a random variable X is denoted by $\mathbb{P}[X]$, and the expected value $\mathbb{E}[X]$. . Where X is a tree T we sometimes emphasise that it has n leaves by writing $\mathbb{P}_n[T]$ for the probability. We represent the number of elements in a set S by $|S|$.

1.3 Rooted Binary Trees

1.3.1 Some Terminology

Evolutionary relationships are often represented by rooted or unrooted *binary (phylogenetic) trees* [53]. In such trees, all nodes of degree 1 are labelled and called *leaves* and all *internal* nodes are unlabelled and of degree 3. Also, in case the tree is rooted, it contains an additional *root node* of degree 2. All trees in this section will be binary. A (tree) *shape* is the unlabelled tree obtained by dropping the labelling of the leaves of a binary phylogenetic tree. A pair of leaves adjacent to a common node is called a *cherry*. Edges

adjacent to a leaf are called *pendant edges*, while all other edges are *internal*. The leaf-set of a labelled tree T is the set of labels the leaves have, and we denote this set by $\mathcal{L}(T)$. For further clarification of these terms see Figure 1.1.

Below any non-leaf node a rooted binary tree T splits into two rooted subtrees a and b, a relationship which we represent by the notation $T = a + b$. If the two subtrees a and b are the same shape then the node is referred to as balanced. The symmetry index of a tree (σ) is equal to the number of balanced nodes it has (see Figure 1.2). If a set of leaves are the only descendants of some internal node then we say that they form a *group* (also called a *clade* or *cluster*).

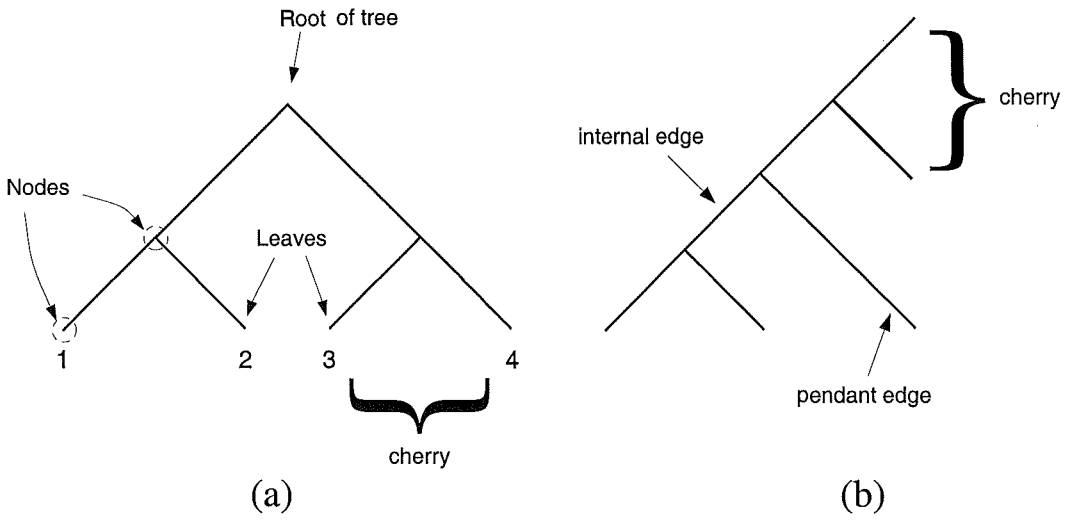


Figure 1.1: Some terminology for trees (a) A labelled rooted tree with 4 leaves. (b) An unrooted tree shape with 5 leaves.

1.3.2 A Dictionary Notation

A non-pictorial symbolic representation for rooted tree shapes is convenient as a shorthand, and for programming purposes. One such non-pictorial representation for rooted tree shapes is the 'dictionary' notation as explicated by Harding [36, p. 59]. In this notation the *ith* shape on n leaves is represented by the symbol n_i. Let the number of tree shapes on n leaves be denoted by $\mathcal{S}(n)$ (see equation (1.1) below), and let r_k be the kth shape on r leaves and s_m be the mth shape on s leaves. We can write n_i recursively in terms of its left subtree r_k and the right subtree s_m as

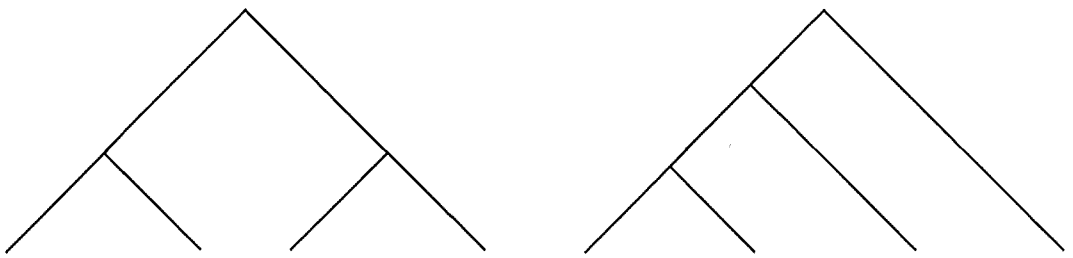


Figure 1.2: Tree balance (a) This is is the *symmetric* tree on 4 leaves (also called the *fully balanced tree*). It has a symmetry index of $\sigma = 3$. (b) The *asymmetric* tree on 4 leaves (also called the *comb tree*, *rooted caterpillar tree*, or *fully unbalanced tree*). It has a symmetry index of $\sigma = 1$.

$$n_i = r_k + s_m \quad (r + s = n,\ r \leq s,\ 1 \leq k \leq \mathcal{S}(r),\ 1 \leq m \leq \mathcal{S}(s)\ ) .$$

For a given value of n the shapes n_i are ordered (1) firstly with respect to the number of leaves on the left subtree (r), (2) secondly with respect to the value of k, (3) lastly with respect to the value of m.

Some examples will help to clarify this notation. For the tree shapes on one, two, and three leaves there is only one tree shape, so the i index is dropped, and they are denoted by the symbols $1, 2, 3$ respectively. The next symbol in this notation is $4_1 = 1+3$ (the fully unbalanced shape on four leaves), followed by $4_2 = 2+2$ (the fully balanced tree shape on four leaves). For $n = 5$ we have three tree shapes represented by the symbols $5_1 = 1 + 4_1$, $5_2 = 1 + 4_2$, and $5_3 = 2 + 3$. For $n = 6$ we have the six tree shapes represented by the symbols $6_1 = 1+5_1$, $6_2 = 1+5_2$, $6_3 = 1+5_3$, $6_4 = 2+4_1$, $6_5 = 2+4_2$, $6_6 = 3+3$. Continuing in this manner we can recursively build up the symbols for all shapes on n leaves. Note, that for a shape with symbol n_i, that $i = 1$ corresponds to the fully unbalanced shape on n leaves, and $i = \mathcal{S}(n)$ corresponds to the most symmetric shape on n leaves.

1.3.3 Enumeration of Rooted Binary Trees

The number of unlabelled rooted binary trees on n leaves is given by [76]

$$\mathcal{S}(n) = \left\lfloor \frac{1}{2} \sum_{k=1}^{n-1} \mathcal{S}(k)\mathcal{S}(n-k) \right\rfloor + E(n) \,, \tag{1.1}$$

where

$$E(n) = \begin{cases} \frac{1}{2}\mathcal{S}(n/2) & \text{if n is even} \\ 0 & \text{if n is odd} \,, \end{cases}$$

and $\mathcal{S}(1) = 1$.

We denote the set of labelled rooted binary trees by $RB(n)$. The number of elements in $RB(n)$ has a simple, explicit formula given by [14]

$$|RB(n)| = (2n-3)!! = (2n-3) \cdot (2n-5) \cdot (2n-7) \ldots 5 \cdot 3 \cdot 1 \,. \tag{1.2}$$

Or equivalently [14], for $n \geq 2$,

$$\frac{(2n-3)!}{2^{n-2}(n-2)!} \,. \tag{1.3}$$

The number of possible labelled trees, for a given tree shape on n leaves, is given by [13]

$$\frac{n!}{2^{\sigma}}, \tag{1.4}$$

where σ is the symmetry index of the shape.

For an example of the usage of the above formulae refer to the diagram below (Figure 1.3).

1.4 The Catalan Numbers

The Catalan numbers and a variety of identities involving them will prove particularly useful when we are dealing with the uniform model (see ahead in Section 1.5.3). In this section we define the Catalan numbers and derive some identities involving them.

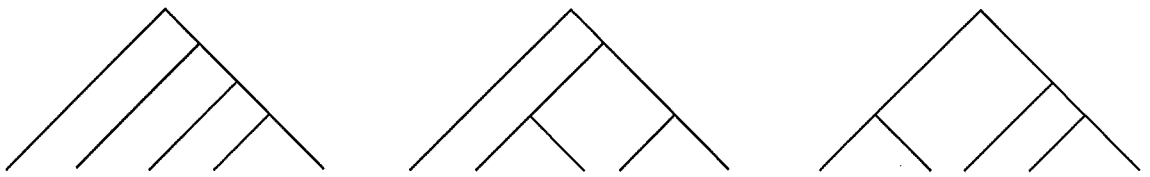


Figure 1.3: The three rooted tree shapes on five leaves. The number of labelled tree shapes on 5 leaves is $7!! = 105$. The tree shape T_1 has $\frac{5!}{2} = 60$ possible labelled trees, while T_2 has 15, and T_3 has 30.

The Catalan numbers (c_n) are frequently defined in a slightly different manner by different authors, the difference simply being whether or not the series of numbers characteristic of the Catalan numbers begins at $n = 0$ or $n = 1$. Here we define the Catalan numbers by

$$c_n = \frac{1}{n-1}\binom{2n-2}{n-2},$$

so that $c_0 = 0,\ c_1 = 1,\ c_2 = 1,\ c_3 = 2,\ c_4 = 5,\ c_5 = 14, \ldots$ are the Catalan numbers [74, p. 172]. They occur as the solutions to several enumerative combinatorics problems such as number of ways to pair up the terms $A_1, A_2, \ldots, A_n$ with parentheses while keeping the original order, the number of plane rooted trees with $n-1$ edges, the number of mountain ranges you can draw using $n-1$ upstrokes and $n-1$ downstrokes, and the number of noncrossing handshakes possible for $n-1$ pairs of people seated around a table [22]. Here we are more concerned with the interpretation of the Catalan numbers as the solutions to the recurrence problem, for $n \geq 2$,

$$c_n = \sum_{k=1}^{n-1} c_k c_{n-k}, \tag{1.5}$$

where $c_0 = 0,\ c_1 = 1$. Associated with this recurrence is the generating function

$$C(x) = \sum_{k=0}^{\infty} c_k x^k = \frac{1}{2} - \frac{1}{2}\sqrt{1-4x}\,, \tag{1.6}$$

which satisfies the functional equation

$$C(x) = x + [C(x)]^2 \,. \tag{1.7}$$

We now collect together and prove some identities which will be useful in later sections.

Lemma 1 *Let c_n be the Catalan number. We have the identities:*

(i) $\sum_{k=1}^{n-1} k c_k c_{n-k} = \frac{n}{2} c_n$.

(ii) $\sum_{k=1}^{n-1} \frac{k}{4^k} c_k = \frac{2n(n-1)c_n}{4^n}$.

Proof. Taking the first derivative of the functional equation (1.7) then multiplying by $\frac{x}{2}$ we obtain

$$\frac{x}{2} C'(x) = \frac{x}{2} + xC(x)C'(x) \,. \tag{1.8}$$

Using (1.6) the lefthand side of (1.8) is equal to

$$\frac{x}{2} + \sum_{n=2}^{\infty} \frac{nc_n}{2} x^n \,, \tag{1.9}$$

and the right hand side is equal to

$$\frac{x}{2} + \sum_{j=0}^{\infty} c_j x^j \sum_{k=0}^{\infty} k c_k x^k = \frac{x}{2} + \sum_{n=2}^{\infty} \left[\sum_{k=1}^{n-1} k c_k c_{n-k} \right] x^n \,. \tag{1.10}$$

Equating the coefficients of x^n of (1.9) and (1.10) gives the first identity.

For the second identity we have from (1.6) that

$$2x^2 C''(x) = \frac{4x}{1-4x} x C'(x) \,,$$

which becomes, after substituting $x/4$ for x,

$$2\left(\frac{x}{4}\right)^2 C''(x/4) = \frac{x}{1-x} \frac{x}{4} C'(x/4) \,. \tag{1.11}$$

For the lefthand side of (1.11) we have

$$\sum_{n=2}^{\infty} \frac{2n(n-1)c_n}{4^n} x^n \,, \tag{1.12}$$

and for the righthand side we have

$$\sum_{j=1}^{\infty} x^j \sum_{k=1}^{\infty} \frac{kc_k}{4^k} x^k = \sum_{n=2}^{\infty} \left[\sum_{k=1}^{n-1} \frac{kc_k}{4^k} \right] x^n \,. \tag{1.13}$$

Equating the coefficients of x^n for (1.12) and (1.13) gives the second identity. □

1.5 Probability Distributions on Trees

1.5.1 Introduction

The set of possible labelled and unlabelled tree shapes is well defined, and in the previous section formulae were given to count the number of elements in these sets. In this section we are concerned with assigning probabilities to trees, both labelled and unlabelled.

Two conditions will be in place for the probability distributions we define. Firstly, since we are dealing with probabilities, then the probabilities for the trees (labelled or unlabelled) must add up to one. For example, there are three unlabelled trees on 5 leaves and their probabilities must add up to one; similarly for the 105 labelled trees on 5 leaves. Secondly, the probability of a labelled tree should be invariant under a different labelling, a property commonly known as *exchangeability.*

We present three probability distributions on trees in this section: Yule, uniform, and comb. Of these three distributions the Yule model is the most important one with regard to the modelling of evolutionary trees, since it is the only one that is an explicit model of the process of speciation. For this reason we will be emphasising results involving the Yule model. The uniform model is useful because it can serve as a model in which trees are on average more imbalanced than those in the Yule model, and analytical results are much easier to obtain with it. The comb model importance is that it represents the extreme of imbalance; only the most imbalanced caterpillar tree shapes are generated in this model.

1.5.2 Yule Model

The *Yule* model (or *Markovian* model) can be defined in many seemingly different ways. One definition, which emphasises the link with modelling speciation, is in terms of the

splitting of pendant edges. In the Yule model on rooted tree shapes each pendant edge has an equal probability of splitting to give birth to two new pendant edges [36]. Or equivalently an edge is added randomly, with a uniform distribution, to a pendant edge at each step (Figure 1.4). This model assumes that speciation is instantaneous, always occurs as bifurcations, is independent across lineages, and that the probability of speciation is the same for all lineages at any given time [44]. Extinction may be incorporated into this model by assuming that the probability of extinction is the same for all lineages, and independent across lineages. If this is the case then a Yule model can still be used, but with a different 'speciation' rate [49, 64].

An alternative definition of the Yule model is in terms of *labelled histories*, normally shortened to simply *histories* [13, 52]. Let the internal nodes of a tree be labelled in time order, with the root having label 1, and subsequent internal nodes the labels $2, 3, \ldots, n-1$ (Figure 1.5). A particular labelled history is identified by the tree shape, leaf labelling, and the time order of the internal nodes. A change in any of these gives a different history (Figure 1.6). In the Yule model each labelled history is equally likely [21, 36, 52].

A final, but important characterisation of the Yule model is in terms of a process in population genetics known as the *coalescent* model [4, 42, 73]. In this model one starts with n objects, then picks two at random to coalesce, giving $n-1$ objects. This process is repeated until there is only a single object left. If this process is reversed, starting with one object to give n objects and marking splits with branches, then it is equivalent to the Yule model. Note that in the coalescent model there is commonly a probability distribution for the times of coalescence, but in the Yule model we ignore this element.

We now deal with finding the probability of trees under the Yule model. Let n be a tree shape on n leaves and $\mathbb{P}[\mathsf{n}]$ be the probability of obtaining this tree shape. This probability may be calculated using a recursive relationship [36], where for a shape $\mathsf{n} = \mathsf{r} + \mathsf{s}$ we have

$$\mathbb{P}_n[\mathsf{n}] = \begin{cases} \frac{2}{n-1}\mathbb{P}[\mathsf{r}]\mathbb{P}[\mathsf{s}] & \mathsf{r} \neq \mathsf{s}\,, \\ \frac{1}{n-1}\mathbb{P}[\mathsf{r}]\mathbb{P}[\mathsf{s}] & \mathsf{r} = \mathsf{s}\,. \end{cases} \tag{1.14}$$

For a labelled tree $t = a + b$ we have the simpler recursion [36]

$$\mathbb{P}[t] = \frac{2}{n-1}\binom{n}{r}^{-1}\mathbb{P}[a]\mathbb{P}[b]\,. \tag{1.15}$$

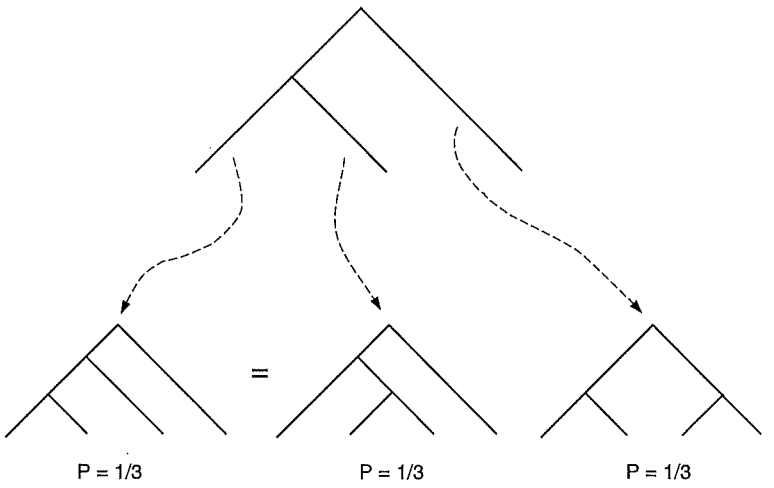


Figure 1.4: The Yule model probabilities for tree shapes on 4 leaves. A shape on 4 leaves is formed by the splitting of one of the pendant edges of the shape on 3 leaves. Each pendant edge has the same probability of splitting, so for the shape on 3 leaves each pendant edge has a probability of 1/3 of splitting. One of the resulting shapes, the symmetric shape on 4 leaves has a probability of 1/3. The other two shapes on 4 leaves are the same (up to rotation about internal nodes), and so the probability of this shape (the rooted caterpillar) is 2/3.

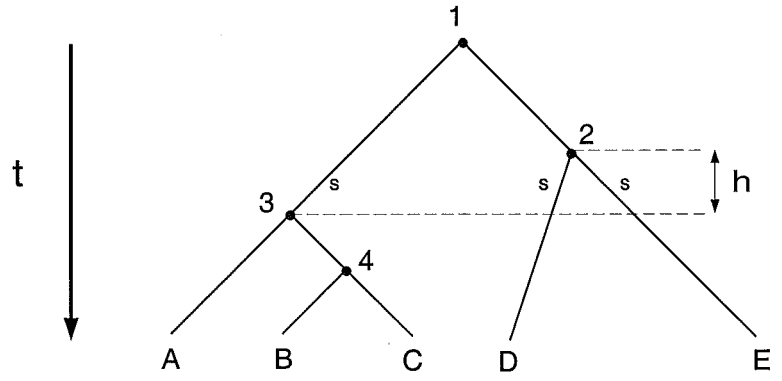


Figure 1.5: Labelled history of a tree. The root of the tree is labelled 1. Successive internodes are labelled 2,3,4. A horizon is the time interval between successive nodes. Here h is the time interval between the nodes labelled 2 and 3. There are 3 internode segments present in this horizon, these being labelled s. In general, between the internodes i-1 and i, there are i internode segments.

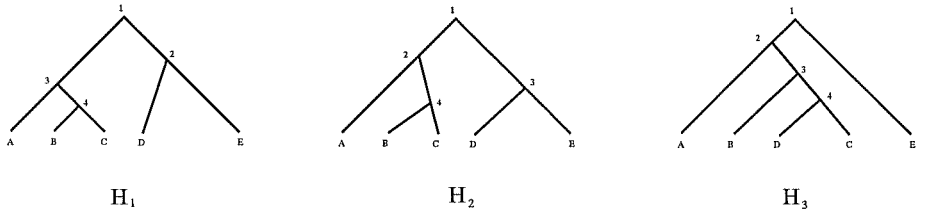


Figure 1.6: Three different labelled histories on 5 leaves. For histories H_1 and H_2 the time order of the speciation events leading to the leaves {A,B,C} and {D,E} is different, though the labelled tree is the same. The history H_3 has a different labelled tree from the histories H_1 and H_2.

Let $\mathbb{P}_n[r,s]$, where $n = r + s$, be the probability that a randomly generated tree shape on n leaves has one subtree with r leaves and the other subtree has s leaves. We have [63]

$$\mathbb{P}[r,s] = \begin{cases} \frac{2}{n-1} & r \neq s\,, \\ \frac{1}{n-1} & r = s\,. \end{cases} \tag{1.16}$$

An alternative way of calculating tree probabilities for the Yule model makes use of its definition in terms of histories. For the set of labelled trees on n leaves the number of possible histories is given by [21]

$$H_n = \frac{n!(n-1)!}{2^{n-1}}\,. \tag{1.17}$$

For a particular labelled tree t on n leaves let the number of histories be denoted by $H_n(t)$. Using this notation then, since each history is equally likely for the Yule model, we have

$$\mathbb{P}[t] = \frac{H_n(t)}{H_n}\,. \tag{1.18}$$

As a simple example, consider the fully unbalanced labelled tree on 4 leaves (see Figure 1.2b). This tree has one history, so the probability of the tree is 1/18. For more complicated trees combinatorial arguments are required in order to find $H_n(t)$ from first principles. Fortunately, general formulae have been found for these calculations [13]. For a labelled tree t on n leaves we have

$$\mathbb{P}[t] = \frac{2^{n-1}}{n!} \prod_{i=1}^{n-1} (a_i - 1)^{-1}\,, \tag{1.19}$$

where a_i is the number of leaves that descend from node i, and the product is over all internal nodes. For the corresponding unlabelled tree t_u on n leaves we have

$$\mathbb{P}[t_u] = \frac{2^{n-1}}{2^{\sigma}} \prod_{i=1}^{n-1} (a_i - 1)^{-1}\,, \tag{1.20}$$

where σ is the symmetry index of the tree.

Let $\delta(v)$ denote the number of internal nodes that are the descendants of v (and including v). Also let $\mathring{T}$ denote the set of internal nodes of a tree T. Then another way of writing equations (1.19) and (1.20) is as

$$\mathbb{P}[t] = \frac{2^{n-1}}{n!} \prod_{v \in \mathring{T}} \delta(v)^{-1} , \tag{1.21}$$

$$\mathbb{P}[t_u] = \frac{2^{n-1}}{2^{\sigma}} \prod_{v \in \mathring{T}} \delta(v)^{-1} . \tag{1.22}$$

To illustrate the use of these formulae consider the labelled tree in Figure 1.5. The probability of this tree is $\frac{24}{5!}[\frac{1}{4} \times \frac{1}{2} \times \frac{1}{1} \times \frac{1}{1}] = \frac{1}{60}$, while for the corresponding unlabelled tree the probability is $\frac{1}{2}$.

1.5.3 Uniform Model

In the *uniform* model (also called the *proportional-to-distinguishable-arrangements* model) on rooted trees, equal probability is assigned to each possible labelled rooted tree on n leaves. For n species there are $(2n-3)!!$ rooted labelled trees, so each tree has a probability of $1/(2n-3)!!$. By counting the number of labelled trees that have a particular shape, using (1.4), the probability of any tree shape may also be calculated (Figure 1.7).

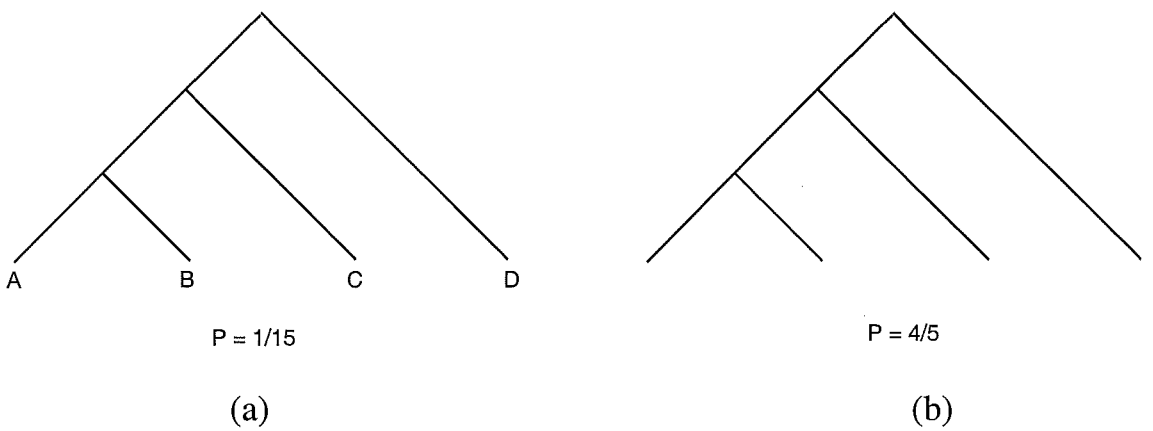


Figure 1.7: The uniform model for 4 species. (a) There are 15 labelled trees for $n = 4$, so any particular labelled tree has probability 1/15. (b) There are 12 labelled trees that have the comb shape, so the comb shape has probability 4/5.

An alternative way of characterising the uniform model on rooted trees is in terms of edge addition. Starting with the rooted tree shape on three leaves, add an edge randomly (with a uniform distribution) to any other edge, allowing for a 'ghost' edge at the root (Figure 1.8a). If an edge is added to the 'ghost' edge then the node that joins them becomes the new root. Repeat this process of edge addition to give a tree shape on n leaves, then assign leaf labels randomly to give a labelled rooted tree. This process is not an explicit model of evolution, though a random sample of n species from a large group of species generated by a conditioned branching process follows a uniform distribution [2, 3]. What the process does model is the tree probability distribution that would occur if the process of tree reconstruction did no better than random selection from the set of possible labelled trees on n leaves. Also since the tree distribution is more imbalanced than the Yule mode, but less than the maximum possible which is the distribution under the comb model, it can be used to 'interpolate' between them.

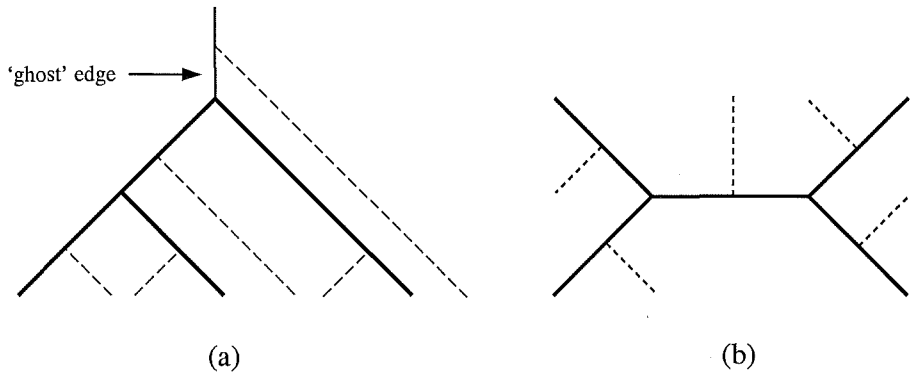


Figure 1.8: The uniform model as a process of edge addition. A tree shape (solid lines) has an edge (dashed lines) uniformly randomly added to one of its edges. (a) The rooted tree shape on three leaves. There are five possible edges where the next edge may be attached. The probability that the next edge will be attached to any particular one of them is 1/5. (b) The unrooted tree shape on four leaves. There are five possible edges where the next edge may be attached. The probability that the next edge will be attached to any particular one of them is 1/5.

In latter problems we will need to know the probability distribution for the number of leaves on the two subtrees of a rooted tree shape. Let $\mathbb{P}_n[r, s]$, where $n = r + s$, be the probability that a randomly generated tree on n leaves has a left subtree with r leaves and

a right subtree with s leaves. We have [63]

$$\mathbb{P}[r,s] = \begin{cases} \frac{2c_r c_s}{c_n}, & r \neq s\ , \\ \frac{c_r^2}{c_n} & r = s\ . \end{cases} \tag{1.23}$$

So far we have been dealing with rooted trees, but the uniform model is also defined on unrooted trees. For n species there are $(2n-5)!!$ unrooted labelled trees, so an unrooted labelled tree has a probability of $1/(2n-5)!!$ [20]. As for the rooted case, this can also be characterised as a process of edge addition (Figure 1.8b), but with the difference that there is no need to incorporate a 'ghost' edge into the process.

1.5.4 Comb Model

The rooted 'comb' shape on n leaves is the most imbalanced tree shape on n leaves. Associated with any internal node of a comb shape are two subtrees, and one of these two subtrees has exactly one leaf.

In the *comb* model the rooted comb shape is assigned a probability of one, while all other possible shapes are assigned a probability of zero (Figure 1.9). From the probability for a shape the probability of a labelled tree of that shape may be calculated, under the condition that all labelled trees of that shape are equally likely. The comb model usefulness is that it represents the extreme of imbalance in tree, and it is not meant as a realistic model for actual evolutionary trees.

1.6 The Empirical Match

1.6.1 Introduction

Early stochastic modelling on phylogenetic trees concentrated on a qualitative comparison between actual phylogenetic trees and those produced in simulation studies [30, 56]. While suggestive, such work suffered from the lack of an explicit measure of the degree of similarity between actual and simulated trees. Later quantitative work, based on simple stochastic models of the process of species formation and extinction, introduced measures of similarity based on the frequency of tree shapes [33, 61, 64] and tree imbalance indices [38, 43, 47, 59].

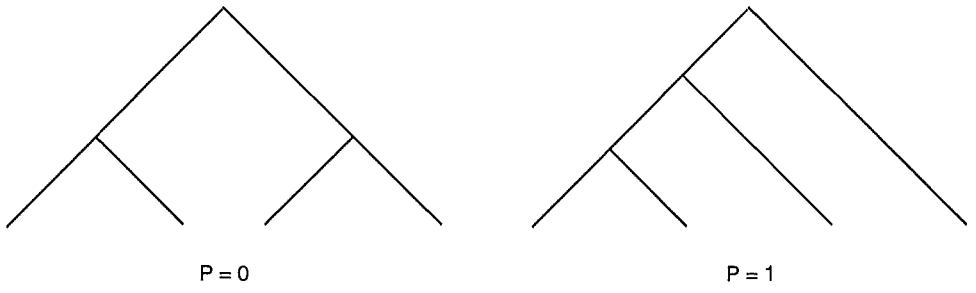


Figure 1.9: The comb model for 4 species. The 'comb' shape is assigned a probability of one, while all other possible shapes are assigned a probability of zero. There are 12 labelled trees that have the comb shape, so the probability for a labelled tree with the comb shape is 1/12.

We have introduced three simple probabilistic models of tree growth: comb, uniform, and Yule. In the next section we review the studies done on the empirical fit between actual (estimated) trees and trees produced in the uniform and Yule models. Of these two models, the Yule model is the only one which is an explicit model of speciation, and thus the one in which there has been the most interest as a null model for speciation. The uniform model is of interest as a null model in which the trees are, on average, more imbalanced than in the Yule model. Also, the uniform model represents the distribution of trees that would occur if tree reconstruction did no better than selecting a labelled tree at random from the set of possible labelled trees. For another review of past work see Mooers [48], where many other aspects of the Yule model are also examined.

Before we begin reviewing past work we need to briefly explain some of the terminology used in the field of tree reconstruction. There are two main schools of classification in use for biological organisms: *phenetic* and *phylogenetic* [57, pp. 355-382]. A phenetic classification groups organisms on the basis of their overall physical similarity, where physical similarity is measured by such characteristics as the shape of bones, size of the organism, skin patterns, the presence of horns or not, and the number of chromosomes. A phylogenetic classification puts organisms into higher or lower taxon groups depending on just how far back their most recent common ancestor was. The most common method for doing phenetic classification is numerical taxonomy, while for phylogenetic classification

cladistics is popular. Despite the methods used for phenetic classification having different goals from the methods used for phylogenetic classification, usually they produce trees that are much the same. Both phenetic and phylogenetic classification schemes are hierarchal, with the hierarchy starting at kingdom and working its way down through phylum (called division in botany), class, order, family, genus, and species. Intermediate levels in the hierarchy are indicated by the prefixes sub, super, etc.

1.6.2 Empirical Fit

The first quantitative study was that of Savage(1983), who compared the frequency of actual tree shapes with those of the uniform and Yule models [61]. Trees with up to 7 leaves and constructed by phenetic and phylogenetic methods were sampled. It was found that for trees on 4,6, and 7 leaves the actual tree shape frequencies did not differ significantly (in a statistical sense) from the predictions of the Yule model, but did from those of the uniform model. For trees on 5 leaves the Yule model null hypothesis was rejected by a χ^2 test at the 5% level, but not at the 2.5% level. Regarding the method of tree construction, it was found that phenetic methods produced more asymmetric trees than phylogenetic methods, in contrast to previous suggestions [16, 62].

Guyer and Slowinski(1991) refined the sampling methodology of Savage by including only trees with species as the terminal branches, and taking a more uniform selection across the higher genera of division Angiospermae, class Insecta, and superclass Tetrapoda [33]. Sampling from trees on five leaves, they found that the frequency of tree shapes were significantly different from the Yule model predictions, but not from the uniform model predictions. This disagrees with the conclusion of Savage, and they suggest that one of the reasons is that including higher taxa as terminal branches (as Savage did) leads to more balanced tree shapes. Additionally, many of the internal nodes of the trees that Guyer and Slowinski used were defined by only a few characters, the effect of this being to move the frequencies towards those of the uniform model. Tree shape frequencies were found not to differ between the three major groups of organisms they sampled.

Heard(1992), in a different approach, investigated the fit between actual and model trees using a corrected form of Colless's tree imbalance index [16, 38]. Heard investigated trees with 4–14 leaves and found that they were significantly more imbalanced than the Yule model trees. Furthermore, this effect did not depend on the data type used (molecular

or morphological), construction method (cladistic or phenetic), or taxon level.

Two taxa are said to be *sister groups* if they are the only descendants of their most recent common ancestor. Guyer and Slowinski(1993) investigated the fit of large trees ($\geq$ 100 leaves) to the uniform and Yule models using a test based on the comparison of sister group sizes. They found their trees to be more imbalanced than the Yule model, but less imbalanced than the uniform model. Their conclusion was independent of the taxon group (division Angiospermae, class Insecta, superclass Tetrapoda).

Mooers(1995) studied the effect of tree incompleteness on tree balance [47]. A tree is *incomplete* if it does not include all the extant species of the clade under consideration. It was found that complete trees were more imbalanced than trees from the Yule model, and incomplete trees even more so. In concurrence with some other studies no significant dependence of balance on construction method (cladistic or phenetic), or taxon level was found [33, 38]. If trees are incomplete due to the random absence of taxa than this should have no effect on tree balance [33]. However, the sampling procedures that systematists use are often not explicitly given, and are unlikely to be random.

1.6.3 Explaining Imbalance

The general consensus of the studies to date is that estimated trees are more imbalanced than those produced by the Yule model, and the imbalance is independent of construction method or taxon level. However, there is still some uncertainty regarding the effect that systematists sampling procedures may have on tree shape and balance.

Taking the imbalance conclusion at face value, one task is to explicate the speciation mechanisms that could give rise to this degree of imbalance. A variety of simple mathematical models have been suggested for this, all based on modifications to the basic assumptions of the Yule model: (1) speciation is instantaneous (2) speciation events are independent (3) the speciation rate is constant across all lineages at a given time.

Losos and Alder(1995) proposed a modification of the Yule model in which the speciation of an organism is not instantaneous, but operates over a finite time during which further speciation events cannot occur [44]. The consequence of the introduction of this 'refractory period' is to make model trees more balanced, thus decreasing even further the empirical fit with actual trees. This concurs with the results of an analytical study we made in which introducing such a refractory period made model trees more balanced, and

even more so for higher speciation rates (see Chapter 9). However, if very long refractory periods are introduced, then the trees produced can be more imbalanced than the Yule model [60]. A related modification is to introduce a refractory period such that if an organism has not speciated up to some finite time ϵ then it never will. Under this modified model trees are more imbalanced compared to the Yule model (Chapter 9). Furthermore, if $t > n\epsilon$, where n is the number of leaves, then the model trees are in fact those from the uniform model [68].

Heard(1996) investigated several models in which the speciation rates were not constant and independent across lineages (key assumptions of the Yule model) [39]. In one model, in which the rate of speciation depended on the value of a "heritable" trait (e.g. body size), it was found that the trees produced were imbalanced compared to the Yule model. In a variety of other models, in which the speciation rate varied across lineages, it was found that the trees produced were also imbalanced. However, as pointed out by Heard, the differences in speciation rate required to produce trees as imbalanced as those in real trees seemed biologically implausible.

1.6.4 Recap

Given that the Yule model is such a simple model, and that it only models the stochastic component of speciation, it would be surprising if actual tree shapes probabilities closely matched the models predictions. That they are at variance seems to borne out by the studies to date.

However, the Yule model is a useful as a null model which can be modified to produce more sophisticated models, such as those already mentioned, that try to explain the tree imbalance observed. Moreover, even in its basic form it is useful as a null model from which patterns in macroevolution need to stand out in order to be considered surprising and worthy of further investigation. A model with a closely analogous role is the Hardy-Weinberg model in genetics for allele frequencies (alleles are different types of a gene). If allele frequencies do not equal those calculated from the Hardy-Weinberg model then this suggests that some evolutionary force is acting on a population.

Chapter 2

Distance Relationships

2.1 Introduction

In this chapter we investigate some 'distance' relationships for rooted trees generated under the Yule and uniform models, where we measure the distance between two vertices by the number of edges separating them. In particular, we are concerned with the distance of a randomly chosen leaf from the root, and with the distance between two randomly chosen leaves. These distances, particularly those under the Yule model, are useful for estimating distances in evolutionary trees where polytomies are present [75].

For the Yule model we find the probability distribution for the distance of a randomly chosen leaf from the root, and from this the mean and standard deviation (Section 2.2.1). We also derive an exact formula for the mean distance between two randomly chosen leaves (Section 2.3.2). For the uniform model we find the mean distance of a randomly chosen leaf from the root (Section 2.2.2), and an exact formula for the mean distance two randomly chosen leaves (Section 2.3.3).

2.2 Distance of a Leaf From the Root

The distance of a leaf from the root is the number of edges along the path from the root to the leaf in question. For example, in Figure 2.1, leaves A and D are both a distance two from the root, while all other leaves are a distance three from the root. From a biological perspective, if we ignore extinction, this distance can be interpreted as the number of speciation events separating the root and a leaf.

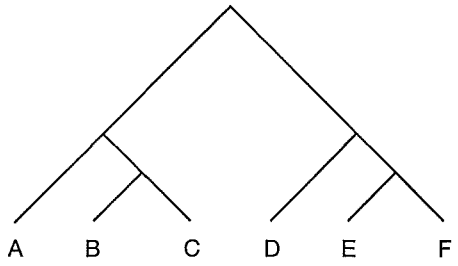


Figure 2.1: A labelled rooted tree with 6 leaves.

2.2.1 The Yule Model

In this section we first find the probability distribution for the distance of a randomly chosen leaf from the root under the Yule model. This was also done in a different context by Lynch [45], but using different methods. In [43] the mean distance of a randomly chosen leaf from the root was derived, and in [4] an asymptotic result for the mean was derived. Here we also derive the variance, a higher order asymptotic result for the mean, and an asymptotic result for the variance.

Theorem 1 *Let P_{n+1}^k be the probability that a randomly chosen leaf from a tree on $n+1$ leaves has distance k from the root. Under the rooted Yule model we have the recursion, for $n \geq 1$ and $k \geq 1$,*

$$P_{n+1}^k = \frac{2}{n+1} P_n^{k-1} + \frac{n-1}{n+1} P_n^k$$

which has the explicit solution, for $n \geq 1$ and $k \geq 1$,

$$P_{n+1}^k = \frac{2^k \left[{n \atop k} \right]}{(n+1)!}$$

where $\left[{n \atop k} \right]$ is the unsigned Stirling number of the first kind [31] .

Proof. By the law of total probability, if $\{B_1, B_2\}$ form a partition of the events that are a precursor to the event A then $\mathbb{P}[A] = \mathbb{P}[A \mid B_1]\mathbb{P}[B_1] + \mathbb{P}[A \mid B_2]\mathbb{P}[B_2]$. Let A be the event that a randomly chosen leaf is distance k from the root for a tree on $n+1$ leaves. Let us call a pair of incident pendant edges in a rooted binary tree a *cherry*. When an edge is added to a pendant edge of a tree on n leaves, this edge, and the edge it was

added to, form a cherry. A randomly chosen leaf from the tree on $n+1$ leaves belongs to one of two mutually exclusive classes: (i) it is a leaf from the new cherry, or (ii) it is one of other leaves. If we let B_1 denote event (i), and B_2 event (ii), then $\mathbb{P}(B_1) = \frac{2}{n+1}$ and $\mathbb{P}(B_2) = \frac{n-1}{n+1}$. For a randomly chosen leaf on a new cherry to be a distance k from the root then the original leaf must have been a distance $k-1$ away, thus $\mathbb{P}[A \mid B_1] = P_n^{k-1}$. If the randomly chosen leaf is not part of a cherry then its distance from the root must have remained unchanged, thus $\mathbb{P}[A \mid B_2] = P_n^k$. Combining all these terms gives the recursion.

The explicit solution to the recursion can be found using generating functions. Denote the ordinary generating function for the probabilities P_n^k by

$$g_n(x) = \sum_{k=0}^{n} P_n^k x^k.$$

From the recursion in the statement of the theorem it follows that

$$g_{n+1}(x) = \frac{2x + (n-1)}{n+1} g_n(x), \qquad \text{where } g_1(x) = 1. \tag{2.1}$$

Solving this recursion gives

$$g_{n+1}(x) = \frac{1}{(n+1)!} \prod_{k=0}^{n-1} (2x + k), \qquad n \geq 1. \tag{2.2}$$

A well known relationship is

$$\prod_{k=0}^{n-1} (x + k) = \sum_{k=1}^{n} \left[{n \atop k} \right] x^k ,$$

where $\left[{n \atop k} \right]$ is the unsigned Stirling number of the first kind [31]. Likewise,

$$\prod_{k=0}^{n-1} (2x + k) = \sum_{k=1}^{n} 2^k \left[{n \atop k} \right] x^k . \tag{2.3}$$

Substituting (2.3) into (2.2) gives

$$g_{n+1}(x) = \frac{1}{(n+1)!} \sum_{k=1}^{n} 2^k \left[{n \atop k} \right] x^k \equiv \sum_{k=0}^{n} P_{n+1}^k x^k . \tag{2.4}$$

Equating the coefficients of x^k gives the formula for P_{n+1}^k. □

Corollary 1 *Let μ_n be the mean distance of a randomly chosen leaf from the root and σ_n^2 the variance. Under the Yule model on rooted trees we have, for $n \geq 2$, that*

$$\mu_n = 2\sum_{k=2}^{n} \frac{1}{k}; \qquad \sigma_n^2 = 2\sum_{k=2}^{n} \frac{1}{k} - 4\sum_{k=2}^{n} \frac{1}{k^2}$$

where $\mu_1 = 0$ and $\sigma_1^2 = 0$. Asymptotically, we have

$$\mu_n - 2\ln n \sim c_1; \qquad \sigma_n^2 - 2\ln n \sim c_2$$

where $c_1 = -2(1-\gamma) \approx -0.846$, $c_2 = 2[1+\gamma] - \frac{2\pi^2}{3} \approx -3.43$, and γ is Euler's constant.

Proof. Noting that $\mu_n = \sum_{k=0}^{n} kP_n^k$ (where $P_n^0 = P_n^n = 0$), the recursion in Theorem 1 implies that

$$\mu_{n+1} = \mu_n + \frac{2}{n+1}, \qquad \text{where } \mu_2 = 1.$$

Solving this recursion gives the explicit solution for the mean. The corresponding asymptotic result follows directly from the relationship

$$\lim_{n\to\infty} \left[\sum_{k=1}^{n} \frac{1}{k} - \ln n \right] = \gamma, \qquad \text{where } \gamma \approx 0.577216 \text{ is Euler's constant [58, p. 15].}$$

For the variance, noting that $\sum_{k=1}^{n} k^2 P_n^k = \sigma_n^2 + \mu_n^2$, and following a similar approach as was used for the mean, leads to the recursion

$$\sigma_{n+1}^2 = \sigma_n^2 + \frac{2}{n+1}\left(1 - \frac{2}{n+1}\right), \qquad \text{where } \sigma_2^2 = 0. \tag{2.5}$$

Solving the recursion gives the explicit solution for the variance. Noting that the first term on the right-hand side of the explicit solution for the variance is the mean, and using the relationship $\sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{\pi^2}{6}$ [58, p. 23] leads to the corresponding asymptotic result for the variance. □

2.2.2 The Uniform Model

An asymptotic result for the mean distance from the root under the uniform model was derived in [4]. Here we derive an exact result for the mean distance from the root, and a higher order asymptotic result.

Theorem 2 *Let ν_n be the mean distance of a randomly chosen leaf from the root for a tree on n leaves. Under the uniform model on rooted trees we have the formula for, $n \geq 1$,*

$$\nu_n = \left(1 - \frac{1}{2n}\right) \frac{2^{2n}}{\binom{2n}{n}} - 1 \,. \tag{2.6}$$

Asymptotically we have

$$\nu_n - \sqrt{\pi n} + \frac{3}{8}\sqrt{\frac{\pi}{n}} \sim -1 \,. \tag{2.7}$$

Proof. In this proof we first find a recursion for ν_n, then solve this recursion to get the exact solution. Let T be a rooted tree on n leaves with a left subtree with k leaves, and a right subtree with $n-k$ leaves. If a leaf is selected at random from T then it will have mean distance from the root given by

$$\frac{k}{n}[\nu_k + 1] + \frac{n-k}{n}[\nu_{n-k} + 1] \,.$$

Conditioning on the size of the subtrees we have

$$\nu_n = \sum_{\substack{r,s \\ r+s=n}} \left[\frac{r}{n}(\nu_r + 1) + \frac{s}{n}(\nu_s + 1)\right] \mathbb{P}_n[r,s]$$

where $\mathbb{P}_n[r,s]$ is the probability that a randomly generated tree shape on n leaves has a left subtree with r leaves and a right subtree with s leaves ($r+s=n$). Using equation (1.23) for $\mathbb{P}_n[r,s]$, and rewriting the summation, we then have the recursion

$$\nu_n = \frac{2}{nc_n} \sum_{k=1}^{n-1} k[\nu_k + 1] c_k c_{n-k} \,,$$

where $c_n = \frac{1}{n-1}\binom{2n-2}{n-2}$ is the Catalan number (see p. 7). Expanding out the summation term, then using the first identity of Lemma 1 gives

$$\nu_n = \frac{2}{nc_n}\sum_{k=1}^{n-1} kc_k c_{n-k}\nu_k + 1 \,. \tag{2.8}$$

Substituting $w_k = kc_k\nu_k$ gives the simpler recursive form

$$w_n = 2\sum_{k=1}^{n-1} c_{n-k}w_k + 1 \,. \tag{2.9}$$

This recursion has the solution

$$w_n = 2^{2n-2} - 2\binom{2n-3}{n-1} \,,$$

so we have

$$\nu_n = \frac{1}{nc_n}\left[2^{2n-2} - 2\binom{2n-3}{n-1}\right] \,.$$

Using the identity $\frac{2}{nc_n}\binom{2n-3}{n-1} = 1$, factoring some terms, and writing out the expression for c_n explicitly gives the stated exact formula. The asymptotic result follows from the relationship $\frac{2^{2n}}{\binom{2n}{n}} = \sqrt{n\pi} + \frac{1}{8}\sqrt{\frac{\pi}{n}} + O(\frac{1}{n^{3/2}})$ [51, p. 1076]. □

2.2.3 Discussion

We have that the mean distance from the root grows as $O(\sqrt{n})$ for the uniform model, and $O(\log(n))$ for the Yule model (and it is easily shown that it grows as $O(n)$ for the comb model). Thus the mean distance grows faster for the uniform model compared to the Yule model. At first this seems counterintuitive if one interprets these models as processes, and considers the addition of the last edge. In the Yule model the last edge can only be added to a pendant edge, while in the uniform model it can be added to any edge including internal edges close to the root. Thus one might expect, since edges can be added close to the root, that a randomly chosen leaf from the uniform model tree would be closer to the root than a randomly chosen leaf from the Yule model tree. In fact this is not the case, because when an edge is added to an internal edge close to the root in the uniform

model tree, the distance of all leaves below this edge increase by one. In the Yule model when an edge is added to a pendant edge the distance of all other leaves from the root remain unchanged (except for the leaf attached to the edge added to). Thus, while the mean distance of the last added edge will be less for the uniform model than for the Yule model, the net effect averaging over all edges is that the mean distance grows faster for the uniform model.

The formula for μ_n is useful for estimating the mean distances of leaves from the root of an unresolved polytomy [75]. For example, for an unresolved polytomy of size four the mean distance of a leaf from the root is $13/6 \approx 2.17$ (under the Yule model). For explicit and asymptotic values of the mean and variance see Table A.1 in Appendix A.1.

2.3 Mean Distance Between Two Leaves

The distance between two leaves is the minimum number of edges that must be transversed in passing from one leaf to the other. For example, in Figure 2.1, the distance E $\leftrightarrow$ F is two, D $\leftrightarrow$ F is three, and A $\leftrightarrow$ D is four. From a biological perspective, if we ignore extinction, then subtracting one from this distance gives the minimum number of speciation events separating two randomly chosen species.

First we prove a lemma which states a recursion for the expected value of the total distance between all $\binom{n}{2}$ pairs of leaves (Section 2.3.1). This recursion is applicable to any distribution on rooted binary trees for which the probability distribution of the left and right subtrees is known. We then look at the particular cases of the Yule model (Section 2.3.2) and uniform model (Section 2.3.3). In both cases we derive exact and asymptotic results for the mean distance between two leaves.

2.3.1 A General Recursion

Let T be a labelled rooted tree on n leaves. Let $f(T)$ be the total of the distance between two leaves, over all $\binom{n}{2}$ pairs of leaves for the tree T. We define D_n to be the expected value of $f(T)$ over all trees T. Let μ_k be the mean distance of a randomly chosen leaf from the root over all trees on k leaves, and $\mathbb{P}_n[r, s]$ the probability that a randomly generated rooted tree has one subtree with number of leaves r and the other subtree has number of leaves $s = n - r$. We have the following recursion for D_n.

Lemma 2

$$D_n = \sum_{all\ shapes} [D_r + D_s + rs(\mu_r + \mu_s + 2)]\mathbb{P}_n[r, s]$$

where the summation is over all tree shapes on n leaves.

Proof. Let T be a random variable representing a labelled tree randomly generated by the Yule process. If the distance between the ith and jth leaves is given by $d(i, j)$ then the total distance between leaves over all $\binom{n}{2}$ pairs of leaves is given by

$$f(T) = \sum_{i<j} d(i, j).$$

Defining $d(i, \rho)$ as the distance of the ith leaf from the root node ρ then the total distance of the leaves on T from the root node ρ is given by

$$h(T) = \sum_{i=1}^{n} d(i, \rho)\ .$$

Let the tree T have the subtrees T_1 and T_2, with leaf sets $\mathcal{L}(T_1)$ and $\mathcal{L}(T_2)$, and roots ρ_1, ρ_2 (Figure 2.2). Then the total distance between the leaves over all $\binom{n}{2}$ pairs is

$$f(T) = f(T_1) + f(T_2) + \sum_{\substack{i\in\mathcal{L}(T_1)\\ j\in\mathcal{L}(T_2)}} \{d(i, \rho_1) + d(j, \rho_2) + 2\}\ .$$

Since

$$\sum_{\substack{i\in\mathcal{L}(T_1)\\ j\in\mathcal{L}(T_2)}} d(i, \rho_1) = sh(T_1), \qquad \sum_{\substack{i\in\mathcal{L}(T_1)\\ j\in\mathcal{L}(T_2)}} d(j, \rho_2) = rh(T_2), \qquad \sum_{\substack{i\in\mathcal{L}(T_1)\\ j\in\mathcal{L}(T_2)}} 2 = 2rs$$

where r,s are the number of leaves on subtrees T_1, T_2 respectively, the random variable $f(T)$ for the total distance becomes

$$f(T) = f(T_1) + f(T_2) + sh(T_1) + rh(T_2) + 2rs.$$

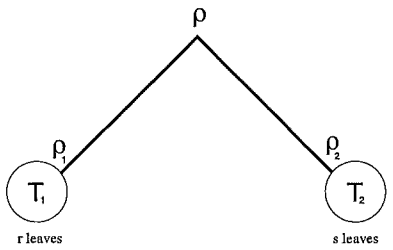


Figure 2.2: Splitting the tree T on n leaves into two subtrees: $T = T_1 + T_2$. The left subtree T_1 has the leaf set $\mathcal{L}(T_1)$ with r leaves, and the root node ρ_1. The right subtree T_2 has the leaf set $\mathcal{L}(T_2)$ with s leaves, and the root node ρ_2.

Since $f(T)$ is a random variable representing the total distance between all $\binom{n}{2}$ pairs of leaves then the mean total distance is

$$D_n = \mathbb{E}[f(T)]. \tag{2.10}$$

Let a tree shape on n leaves be made up of two subtree shapes with r and s leaves $(r + s = n)$ then we have

$$D_n = \sum_{r+s=n} \mathbb{E}[f(T)|r, s]\mathbb{P}_n[r, s].$$

The expected value term is

$$\begin{aligned} E[f(T)|r, s] &= \mathbb{E}[f(T_1) + f(T_2) + sh(T_1) + rh(T_2) + 2rs] \\ &= D_r + D_s + rs\mu_r + rs\mu_s + 2rs, \end{aligned}$$

and substituting this in gives the required recursion. □

2.3.2 The Yule Model

We now solve the recursion in Lemma 2 for the case where the probability distribution of the subtrees is that of the Yule model. We obtain an exact solution for the mean distance between two randomly chosen leaves (d_n). For values of d_n see Table A.3 in Appendix A.2.

Theorem 3 *Under the rooted Yule model for labelled trees on n leaves we have, for $n \geq 2$,*

$$d_n = \frac{2(n+1)}{(n-1)}\mu_n - 4\,, \tag{2.11}$$

where $\mu_n = 2\sum_{j=2}^{n} \frac{1}{j}$. Asymptotically, we have

$$d_n - 4\ln n \sim c\,, \tag{2.12}$$

where $c = 4\gamma - 8 \approx -5.69$, and γ is Euler's constant.

Proof. We first find a recursion for d_n, then find the explicit solution for the recursion. Substituting (1.16) for $\mathbb{P}_n[r,s]$ in Lemma 2 then we for the expected total distance

$$D_n = \frac{1}{n-1} \sum_{\substack{r,s \\ r+s=n}} [D_r + D_s + rs(\mu_r + \mu_s + 2)]\,,$$

where $\mu_n = 2\sum_{j=2}^{n} \frac{1}{j}$ (as in Corollary 1). Expanding the recursion out further then we have

$$\begin{aligned} D_n &= \frac{1}{n-1} \sum_{k=1}^{n-1} [D_k + D_{n-k} + k(n-k)(\mu_k + \mu_{n-k} + 2)] \\ &= \frac{2}{n-1} \sum_{k=1}^{n-1} D_k + \frac{2}{n-1} \sum_{k=1}^{n-1} k(n-k)\mu_{n-k} + \frac{1}{3}n(n+1)\,. \end{aligned}$$

Letting $f(n) = \frac{2}{n-1}\sum_{k=1}^{n-1} k(n-k)\mu_{n-k} + \frac{1}{3}n(n+1)$, the recursion for the total distance can be rewritten as

$$D_n = \frac{2}{n-1} \sum_{k=1}^{n-1} D_k + f(n)\,.$$

Substituting for the term D_n, in the expression for D_{n+1}, leads to an alternative form for the recursion

$$D_{n+1} = \left(\frac{n+1}{n}\right) D_n + g(n)\,, \tag{2.13}$$

where $g(n) = f(n+1) - \frac{n-1}{n} f(n) = \frac{2}{n} \sum_{k=1}^{n} k\mu_k + n + 1$. We now solve this recursion. The explicit solution for (2.13), with C an arbitrary constant, is [74, page 233]

$$\begin{aligned} D_n &= \left(\prod_{k=2}^{n-1} \frac{k+1}{k} \right) C + \sum_{m=2}^{n-2} \left(\prod_{k=m+1}^{n-1} \frac{k+1}{k} \right) g(m) + g(n-1) \\ &= n + n \sum_{m=2}^{n-2} \frac{g(m)}{m+1} + g(n-1) \qquad \text{where } C = 2 \text{ since } D_3 = 8 \ . \end{aligned} \tag{2.14}$$

An important term in the explicit solution is $\sum_{k=1}^{n} k\mu_k$. To evaluate this, a useful identity [51, p. 1078] known as summation by parts, is

$$\sum_{j=1}^{n} a_j b_j = \sum_{k=1}^{n-1} A(k)(b_k - b_{k+1}) + A(n) b_n \ , \tag{2.15}$$

where $A(k) = \sum_{j=1}^{k} a_j$. Letting $a_j = j$ and $b_j = \mu_j$ gives

$$\sum_{k=1}^{n} k\mu_k = -\frac{1}{2}(n-1)n + \frac{1}{2} n(n+1)\mu_n \ .$$

Using this expression gives for the last term of (2.14)

$$g(n-1) = -(n-2) + n\mu_{n-1} + n \ . \tag{2.16}$$

Using (2.15) and (2.16), then the second term of (2.14) is

$$n \sum_{m=2}^{n-2} \frac{g(m)}{m+1} = n(n+1)\mu_n - n\mu_{n-1} - 2n^2 + n - 2 \ . \tag{2.17}$$

Combining all the terms together gives the explicit solution for the total distance, which is related to the mean distance by $d_n = D_n / \binom{n}{2}$, thus giving the explicit solution for the mean distance between two leaves. The asymptotic result follows from the asymptotic expression for μ_n. □

2.3.3 The Uniform Model

In [69] an exact formula for the mean distance between two leaves for unrooted trees satisfying the uniform model was derived. Here we derive an exact formula for the mean distance $\mathtt{d_n}$ between two leaves for rooted trees, and an asymptotic result.

Theorem 4 *Under the uniform model for rooted labelled trees on n leaves we have, for $n \geq 1$,*

$$\mathtt{d_n} = \left(1 - \frac{1}{2n}\right) \frac{2^{2n}}{\binom{2n}{n}} . \tag{2.18}$$

Asymptotically we have

$$\mathtt{d_n} - \sqrt{\pi n} \sim -\frac{3}{8}\sqrt{\frac{\pi}{n}} . \tag{2.19}$$

Proof. We prove the result by deriving a recursion for $\mathtt{d_n}$, then solving this recursion. Starting with the general recursion for the total distance (Lemma 2), then using (1.23) to substitute for $\mathbb{P}_n[r, s]$, we obtain the following recursion

$$D_n = \frac{1}{c_n} \sum_{\substack{r,s \\ r+s=n}} [D_r + D_s + rs(\nu_r + \nu_s + 2)]c_r c_s .$$

Substituting $D_n = \binom{n}{2}\mathtt{d_n}$, then separating out the terms in the square brackets, leads to

$$\mathtt{d_n} = \frac{2}{n(n-1)c_n} \sum_{k=1}^{n-1} k(k-1)c_k \mathtt{d_k} + \frac{4}{n(n-1)c_n} \sum_{k=1}^{n-1} k(n-k)(\nu_k + 1)c_k c_{n-k} .$$

For the last term, if we substitute for $\nu_k + 1$, rewrite $c_k = \binom{2k}{k}/(2(2k-1))$, then apply the second identity in Lemma 1 we get the value two. Therefore we have

$$\mathtt{d_n} = \frac{2}{n(n-1)c_n} \sum_{k=1}^{n-1} k(k-1)c_k \mathtt{d_k} + 2 .$$

Substituting $x_n = n(n-1)c_n \mathtt{d_n}$, a simpler form for the recursion,

$$x_n = 2\sum_{k=1}^{n-1} c_{n-k} x_k + 2n(n-1)c_n ,$$

is obtained.

This recursion has the solution $x_n = (n-1)4^{n-1}$ so we have $\mathtt{d_n} = \frac{4^{n-1}}{nc_n}$. Using the explicit form for c_n and rearranging gives the stated explicit form for $\mathtt{d_n}$. The asymptotic result follows directly from the asymptotic result in Theorem 2. □

2.3.4 Discussion

For the Yule model the mean distance between two leaves grows as $O(\log(n))$ while for the uniform model the mean distance grows as $O(\sqrt{n})$. These growth rates exemplify a general point made by Aldous [2], that almost all examples of random trees fall into two categories with regard to the mean distance between two leaves: those for which the mean distance grows as $O(\log(n))$, and those that grow as $O(\sqrt{n})$. The same point could be made for rooted trees with regard to the mean distance from the root, again exemplified by the Yule and uniform models.

Let $d_{RM}(n)$ be the mean distance separating the root node and the most recent common ancestor (MRCA) of two randomly chosen leaves. This quantity is closely related to the mean distance between two leaves by the expression $d_{RM}(n) = \mu_n - \frac{1}{2}d_n$, where μ_n is the mean distance of a leaf from the root and d_n is the mean distance between two leaves (see Steel and McKenzie in [68]). Substituting in the appropriate expressions for the Yule model we obtain $d_{RM}(n) = (1 - \frac{n+1}{n-1})\mu_n + 2 \sim 2$, where the asymptotic value of 2 is an upper bound. Doing likewise for the uniform model we obtain $d_{RM}(n) = (1 - \frac{1}{2n})\frac{2^{2n-1}}{\binom{2n}{n}} - 1 \sim \frac{1}{2}\sqrt{\pi n}$. Thus for the uniform model the mean distance between the root node and the MRCA of two randomly chosen leaves increases with the size of the tree, while in contrast for the Yule model this mean distance has an upper bound of two.

Chapter 3

Distribution Of Cherries For Two Models of Trees

3.1 Introduction

In this chapter we consider a simple and easily computed statistic for tree shape – namely the number of pairs of leaves that are adjacent to a common node. Such a pair of leaves we will call a *cherry* (see Figure 1.1). Firstly, *extended Polya urn* models are explained (Section 3.2). We then derive asymptotic normality results for the number of cherries in the Yule and uniform models, as well as exact results for the mean and variance (Section 3.3). These results are used to develop statistical tests for the Yule and uniform null hypotheses, and the power of these tests is also calculated using the other model as an alternative hypothesis (Section 3.4). The use of the statistical tests is illustrated with an application to a 34-species tree. We then consider an extension of the urn model for cherries, where we look at the asymptotic distribution of *triplets* - where a triplet consists of a cherry and a leaf that are adjacent to a common node (Section 3.5).

3.2 Extended Polya Urn (EPU) Models

In this section we review a recent central limit theorem concerning a general type of urn model, which will be useful for describing the asymptotic distribution of cherries.

Suppose an urn contains p types of balls. If a ball of the ith type ($i \in \{1, \ldots, p\}$) is drawn from the urn then it is returned, along with A_{ij} balls of the jth type. A_{ij} can be

negative, this corresponding to the removal of balls from the urn. Models with $A_{ii} \geq 0$ (and commonly $A_{ij} \geq 0$) are referred to as *generalized Polya urn* (GPU) models [7, 8]. Allowing for A_{ii} to be negative, but requiring the number of balls returned each time to be a positive constant, defines the class of *extended Polya urn* (EPU) models [9, 66].

For both classes of urn models a number of asymptotic normality results exist, but in this paper we need only consider some specific asymptotic results for the EPU model, as follows [9, 66].

Theorem 5 *[9, 66] Let $A = [A_{ij}]$ be the generating matrix for an EPU model, with principal eigenvalue λ_1. Let v be the left eigenvector of A corresponding to λ_1, where the entries v_i add up to one. Also let Z_{in} denote the number of balls of type i in the urn after n draws, where $i = 1, 2, \ldots, p$. For $p = 2$ suppose that:*

(i) A has constant row sums, where the constant is positive,

(ii) λ_1 is positive, simple, and has a strictly positive left eigenvector v,

(iii) $2\lambda < \lambda_1$ for the non-principal eigenvalue λ;

then $n^{-1/2}(Z_{1n} - n\lambda_1 v_1)$ has asymptotically a normal distribution with mean of zero.

Furthermore, for $p > 2$, suppose in addition:

(iv) $2Re(\lambda) < \lambda_1$ for all non-principal eigenvalues λ,

(v) all complex eigenvalues are simple, and no two distinct complex eigenvalues have the same real part, except for conjugate pairs,

(vi) all eigenvectors are linearly independent;

then $n^{-1/2}(Z_{1n} - n\lambda_1 v_1, Z_{2n} - n\lambda_1 v_2, \ldots, Z_{(p-1)n} - n\lambda_1 v_{(p-1)})$ has asymptotically a joint normal distribution with mean of zero.

3.3 Probability Distribution for Cherries

In this section we find the probability distribution for cherries under the rooted Yule model, and the unrooted uniform model. Throughout we use C_n as a random variable for the number of cherries on a tree on n leaves, and define $P_n^k = \mathbb{P}[C_n = k]$.

3.3.1 Yule Model

In the following lemma we state a recursion for the probability distribution for the number of cherries. This lemma was implicit in Steel and Penny [69], but without formal proof.

Lemma 3 *Let P_n^k be the probability that a binary tree on n leaves has k cherries. Under the Yule model (rooted or unrooted) we have the recursion, for $n \geq 4$,*

$$P_{n+1}^k = \left[1 - \frac{2(k-1)}{n}\right] P_n^{k-1} + \frac{2k}{n} P_n^k. \tag{3.1}$$

For the rooted Yule model we have the initial values $P_4^1 = 2/3, P_4^2 = 1/3$, while for the unrooted Yule model we have $P_4^0 = 0, P_4^2 = 1$.

Proof. By the law of total probability, if $\{B_1, B_2\}$ form a partition of the events that are a precursor to the event A then $\mathbb{P}[A] = \mathbb{P}[A \mid B_1]\mathbb{P}[B_1] + \mathbb{P}[A \mid B_2]\mathbb{P}[B_2]$. Let A be the event that a tree on $n+1$ leaves has k cherries. Let B_1, B_2 be the events that the tree on n leaves has $k, k-1$ cherries respectively. We have $\mathbb{P}[B_1] = P_n^k$ and $\mathbb{P}[B_2] = P_n^{k-1}$. If an edge is attached to a tree on n leaves then it can either (i) attach to a pendant edge that is part of a cherry, leaving the number of cherries unchanged or (ii) attach to some other pendant edge, increasing the number of cherries by one. Hence we have $\mathbb{P}[A|B_1] = \frac{2k}{n}$ and $\mathbb{P}[A|B_2] = 1 - \frac{2(k-1)}{n}$. Combining all the terms gives the recursion. The initial values follow from the probabilities of the rooted and unrooted tree shapes on four leaves. □

Using this lemma we find the mean and standard deviation for the number of cherries under the Yule model, as follows.

Theorem 6 *[69] Let μ_n be the mean number of cherries for a rooted binary tree on n leaves, and σ_n^2 be the variance for the number of cherries. Under the Yule distribution we have the recursions, for $n \geq 2$:*

$$\mu_{n+1} = 1 + \mu_n(1 - \frac{2}{n}); \quad \sigma_{n+1}^2 = \sigma_n^2(1 - \frac{4}{n}) + \frac{2}{n}\mu_n(1 - \frac{2}{n}\mu_n)$$

which may be solved exactly to give

$$\mu_n = \frac{n}{3} \ (n \geq 3); \quad \sigma_n^2 = \frac{2n}{45} \ (n \geq 5).$$

Proof. The proof relies on the following recursion for the probability generating function for C_n under the Yule model. Let $P_n(x) := \sum_{k\geq 1} P_n^k x^k$. Then, from Lemma 3 we have the recursion, for $n \geq 4$:

$$P_{n+1}(x) = xP_n(x) + \frac{2x}{n}(1-x)\frac{d}{dx}P_n(x)\ . \tag{3.2}$$

From this the recursions for the mean and variance follow directly by the usual techniques (noting that $\mu_n = \frac{d}{dx}P_n(x)|_{x=1}$ and $\sigma_n^2 = \frac{d^2}{dx^2}P_n(x)|_{x=1} + \mu_n - \mu_n^2$). The stated explicit formulas for the mean and variance can then be verified by induction on n. □

The asymptotic probability distribution for the number of cherries in the Yule model may be found by realizing the process of cherry formation in an EPU model. Let the pendant edges that are part of a cherry be represented by black balls, and the rest of the pendant edges be represented by white balls. The number of cherries is then half the number of black balls, and the total number of pendant edges is equal to the number of black and white balls.

The following urn scheme generates a probability distribution for the number of black balls that is equal to the probability distribution for the number of pendant edges that are part of some cherry. Starting with a rooted binary tree with two leaves, put two black balls into an empty urn. Select a ball at random from the urn, then return it. If the ball selected was black then put in a white ball. If the ball selected was white then put in two black balls, and take out a white ball. Repeat the process of random selection, and the addition or removal of balls $n-2$ times, until there are n balls in the urn.

The generating matrix for this urn scheme is

$$A = \begin{pmatrix} 0 & 1 \\ 2 & -1 \end{pmatrix}$$

where balls of type one are black, and those of type two are white.

The eigenvalues of A are the principal eigenvalue $\lambda_1 = 1$, and $\lambda_2 = -2$. For the principal eigenvalue the left eigenvector, for which the entries add up to one, is $v = (2/3\ 1/3)^T$. The conditions for the EPU model asymptotic results to apply in Theorem 5 are clearly satisfied so

$$\frac{1}{\sqrt{n}}(Z_{1n} - 2n/3) \to \mathcal{N}(0, c)$$

where $\mathcal{N}(\mu, \sigma^2)$ is a normal distribution with mean μ and variance σ^2.

Substituting $Z_{1n} = 2C_n$ gives

$$\frac{C_n - n/3}{\sqrt{cn/2}} \to \mathcal{N}(0, 1).$$

Using Theorem 6 the value of the constant c can be identified giving the following result.

Corollary 2 *For the Yule model on rooted trees*

$$\frac{C_n - n/3}{\sqrt{2n/45}} \to \mathcal{N}(0, 1).$$

3.3.2 Uniform Model

Theorem 7 *[69, 40, 71] Let μ_n be the mean for C_n for an unrooted binary tree on n leaves, and σ_n^2 be the variance for C_n. Under the uniform model, for $n \geq 4$,*

(a)

$$\mathbb{P}[C_n = k] = \frac{n!(n-2)!(n-4)!2^{n-2k}}{(n-2k)!(2n-4)!k!(k-2)!}, \quad k \geq 2$$

(b)

$$\mu_n = \frac{n(n-1)}{2(2n-5)} \sim \frac{n}{4}; \quad \sigma_n^2 = \frac{n(n-1)(n-4)(n-5)}{2(2n-5)^2(2n-7)} \sim \frac{n}{16}.$$

Proof. Part (a) is due to Hendy and Penny [40], while the first part of (b) appears in Steel and Penny [69]. For the second (variance) part of (b) we note from Steel [71] that the pth cumulative moment of C_n is $\frac{(2(n-s)-5)!!n!}{2^s(2n-5)!!(n-2s)!}$ (see equation (1.2) for the !! notation) and thus,

$$\sigma_n^2 = \frac{n(n-1)(n-2)(n-3)}{2^2(2n-5)(2n-7)} + \mu_n - \mu_n^2.$$

Rearranging this last equation gives the result. □

The asymptotic probability distribution for the number of cherries in the uniform model may be found by an extension of the urn scheme for the Yule model. As before, let pendant edges that are part of cherries be represented by black balls and the other

pendant edges by white balls. In addition, let the internal edges be represented by red balls.

The uniform model, with regard to the number of cherries, is equivalent to the following urn scheme. Starting with an unrooted binary tree with four leaves, put four black balls and one red ball into an empty urn. Select a ball at random from the urn, then return it. If the ball selected was black then put in a white ball and a red ball. If the ball selected was white then put in two black balls, take out a white ball, and put in a red ball. If the ball selected was red then put in a white ball and a red ball. Repeat the process of random selection, and the addition or removal of balls $n-4$ times, until there are n balls in the urn.

The generating matrix for this urn scheme is

$$A = \begin{pmatrix} 0 & 1 & 1 \\ 2 & -1 & 1 \\ 0 & 1 & 1 \end{pmatrix}$$

where balls of type one are black, type two are white, and type three are red.

The eigenvalues of A are $-2, 0$ and the principal eigenvalue $\lambda_1 = 2$. For the principal eigenvalue the left eigenvector of A, for which the entries add up to one, is $v = (1/4\ 1/4\ 1/2)^T$. The conditions for the EPU model asymptotic results to apply in Theorem 5 are clearly satisfied so

$$\frac{1}{\sqrt{n}}(Z_{1n} - n/2) \to \mathcal{N}(0, c).$$

Substituting $Z_{1n} = 2C_n$ gives

$$\frac{C_n - n/4}{\sqrt{cn/4}} \to \mathcal{N}(0, 1).$$

Using Theorem 7 the value of the constant c can be identified, giving the following result.

Corollary 3 *For the uniform model on unrooted trees*

$$\frac{C_n - n/4}{\sqrt{n/16}} \to \mathcal{N}(0, 1).$$

3.3.3 Rooted and Unrooted Trees

The Yule and uniform models, as stochastic processes involving random edge addition, can apply to both rooted and unrooted trees. In the Yule model an edge is added uniformly and randomly to a pendant edge, while in the uniform model an edge is added uniformly and randomly to any edge (allowing a 'ghost' edge at the root in the rooted case). For the process of generating leaf-labelled trees two possible schemes are to add taxa in either fixed order or uniformly randomly. For the Yule model the taxa must be added uniformly randomly in order to generate the correct probability distribution on leaf-labelled trees, while for the uniform model either scheme can be used.

We have so far only considered the cherry distribution for the Yule process on rooted trees and the uniform process on unrooted trees. How does the cherry distribution change if rooted trees are "unrooted" by suppressing the root, or if unrooted trees are "rooted" by the introduction of a root? A related question is what is the cherry distribution for the Yule process on *unrooted* trees, and the uniform process on *rooted* trees?

Consider firstly the unrooting of a rooted tree. Let T be a rooted tree on n leaves, with the number of cherries given by the random variable C_n. If the (degree 2) root of T is suppressed then the number of cherries (C_n^*) either remains the same, or increases by one; the latter occurs precisely when the tree shape has the generic shape shown in Figure 3.1. Let $D_n = C_n^* - C_n \in \{0, 1\}$. We have the following lemma for D_n.

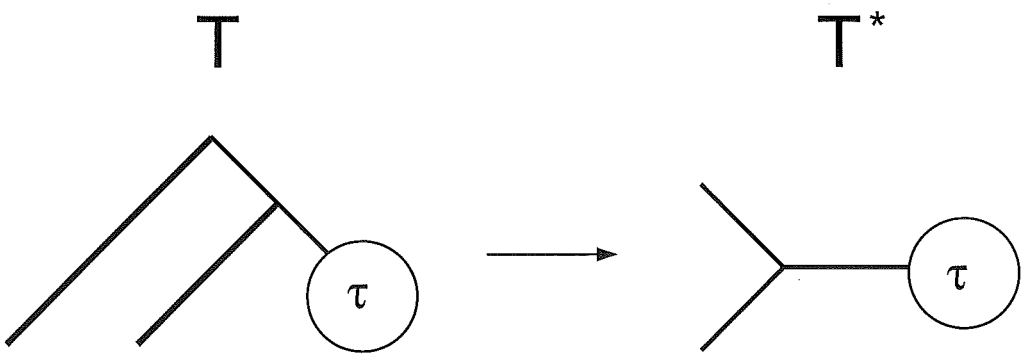


Figure 3.1: Rooted binary tree T for which the associated unrooted tree T^* has one more cherry. The tree T has n leaves, and the subtree τ has $n-2$ leaves.

Lemma 4

(i) For the Yule model on rooted trees

$$\lim_{n\to\infty} \mathbb{P}[D_n = 1] = 0 \,.$$

(ii) For the uniform model on rooted trees

$$\lim_{n\to\infty} \mathbb{P}[D_n = 1] = 1/4 \,.$$

Proof. D_n equals one precisely when the rooted tree shape is as shown in Figure 3.1. For the Yule model, two applications of the recursive formula for tree probabilities (1.14) gives

$$\mathbb{P}[D_n = 1] = \frac{4}{(n-1)(n-2)} \sum_{\tau} \mathbb{P}[\tau] = \frac{4}{(n-1)(n-2)} \,,$$

where the summation is over all subtree shapes τ on $n-2$ leaves. Taking the limit gives the required result. For part (b), there are $(2n-3)!!$ possible leaf-labelled rooted trees on n leaves. Of these $n(n-1)(2n-7)!!$ have the shape shown in Figure 3.1. Therefore

$$\mathbb{P}[D_n = 1] = \frac{n(n-1)(2n-7)!}{(2n-3)!!} = \frac{n(n-1)}{(2n-3)(2n-5)} \,,$$

and taking the limit gives the required result for the uniform model. □

If an unrooted tree is rooted by subdividing some edge, then the number of cherries either remains the same or decreases by one; the latter occurs precisely when the tree is rooted on an edge that is part of a cherry. Since rooting or unrooting a tree changes the number of cherries by a maximum of one the asymptotic probability distribution for the number of cherries will remain unchanged.

Furthermore, the Yule and uniform processes can apply on both rooted and unrooted trees. For both processes the generating matrix for the corresponding EPU model is the same in the rooted and unrooted cases. Therefore, the asymptotic probability distribution for the number of cherries is the same for the rooted and unrooted versions of the Yule and uniform processes.

3.4 Statistical Tests

3.4.1 The Yule Model Null Hypothesis

The Yule model can be used as a simple null hypothesis to explore patterns in phylogenetic trees. A simple two-tailed test of the Yule null hypothesis, for a given tree, can be made based on the number of cherries in the tree. If the number of cherries is below some lower critical value, or above some upper critical value, then the Yule null hypothesis is rejected.

For small n, the recursive formula of Lemma 3 may be used to calculate the rejection limits (Figure 3.2). For larger values of n ($n \gtrsim 20$) a normal approximation is valid. In this case, based on Corollary 2, the rejection region for a two-sided test at the α level is given by

$$C_n < \frac{n}{3} - Z_{\frac{\alpha}{2}}\sqrt{\frac{2n}{45}} \qquad \text{and} \qquad C_n > \frac{n}{3} + Z_{\frac{\alpha}{2}}\sqrt{\frac{2n}{45}}.$$

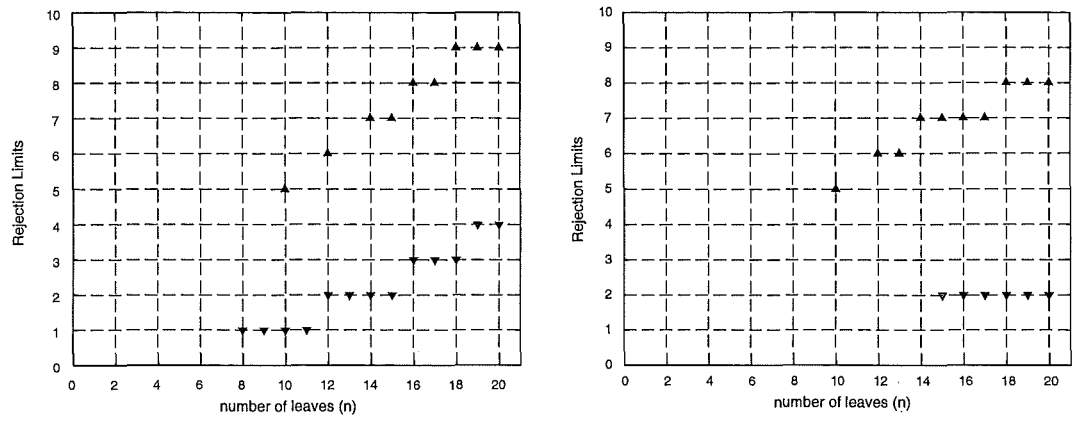
The lower and upper critical values for rejection at an $\alpha = 0.05$ level are shown in Figure 3.3a. If the Yule model is rejected then this implies that one or more of the assumptions upon which it is based is invalid. Often it is assumed that the assumption of equal probability of speciation is the invalid assumption, but this need not be the case [44].

3.4.2 Uniform Model Null Hypothesis

In the uniform model equal probability is assigned to each possible leaf-labelled binary tree on n leaves. Thus the uniform model distribution may be used to model the frequency of outcomes that would occur if the process of tree reconstruction did no better than random selection from the set of possible trees on n leaves. A test of the uniform model null hypothesis may be constructed based on the number of cherries in a tree. For small n the probability distribution given in Theorem 7 may be used to calculate the rejection limits (Figure 3.2b). For larger n ($n \gtrsim 20$) a analysis similar to that for the Yule model, but based on Corollary 3, gives as the rejection region:

$$C_n < \frac{n}{4} - Z_{\frac{\alpha}{2}}\sqrt{\frac{n}{16}} \qquad \text{and} \qquad C_n > \frac{n}{4} + Z_{\frac{\alpha}{2}}\sqrt{\frac{n}{16}}.$$

The lower and upper critical values for rejection at an $\alpha = 0.05$ level are shown in Figure 3.3b.



(a) Yule model (b) Uniform model

Figure 3.2: Rejection limits for small n of the Yule and uniform null hypotheses at the $\alpha = 0.05$ level. Lower limits (▼) and upper limits (▲) were calculated from the exact probability distribution for the number of cherries. Where no triangle is shown the rejection limit does not exist.

3.4.3 Power of Tests

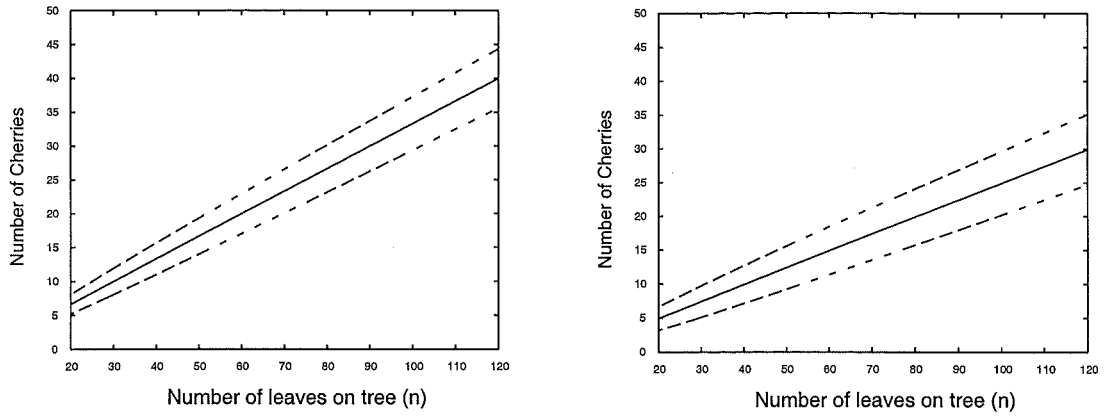
The *power* of a test is the probability that the null hypothesis will be rejected given that the alternative hypothesis is true. Calculating the power of the test for the Yule null hypothesis against the uniform model alternative hypothesis gives

$$\text{power}(n) = \mathbb{P}\left[Z < \sqrt{n}/3 - r_1\right] + \mathbb{P}\left[Z > \sqrt{n}/3 + r_1\right], \qquad r_1 = 4\sqrt{\frac{2}{45}}Z_{\frac{\alpha}{2}}.$$

Similarly, the power of the test for the uniform model null hypothesis may be calculated against the alternative hypothesis that the Yule model is true to give

$$\text{power}(n) = \mathbb{P}\left[Z < -\frac{1}{4}\sqrt{\frac{5}{2}}\sqrt{n} - r_2\right] + \mathbb{P}\left[Z > -\frac{1}{4}\sqrt{\frac{5}{2}}\sqrt{n} + r_2\right], \qquad r_2 = \frac{1}{4}\sqrt{\frac{45}{2}}Z_{\frac{\alpha}{2}}.$$

Plotting the power, as a function of n, shows that in both cases the number of leaves must exceed 80 before the power of the tests rises above 0.9 (Figure 3.4). So, unless one is dealing with large trees, the tests lacks the power to distinguish between the two models.



(a) Yule model (b) Uniform model

Figure 3.3: Rejection limits for large n of the Yule and uniform null hypotheses at the $\alpha = 0.05$ level. The solid line represents the mean number of cherries, while the dashed lines are the lower and upper limits for rejection of the null hypotheses. The rejection limits are based upon a normal approximation which is valid for n $\gtrsim$ 20.

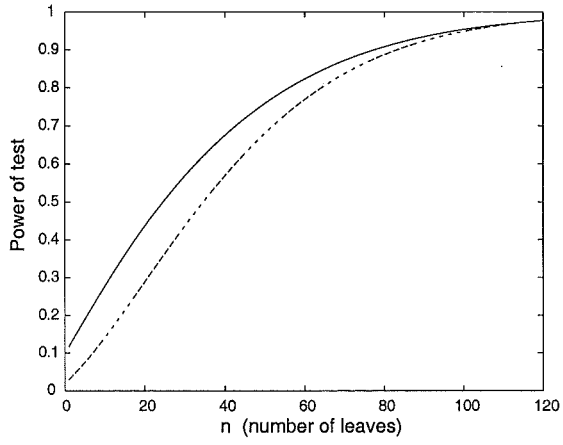


Figure 3.4: The power of the tests for the Yule and uniform models. The solid line is the power of the test for the Yule model null hypothesis against the uniform model alternative. The dashed line is the power of the test for the uniform model null hypothesis against the Yule model alternative.

3.4.4 An Example

Figure 1 in [37] is a rooted phylogenetic tree for 34 species of *eureptantic nemerteans* (ribbon worms). This tree (rooted or unrooted) has 7 cherries. For the Yule model null hypothesis test at the $\alpha = 0.05$ level the lower rejection limit is 8 cherries or less, and the upper rejection limit is 15 cherries or more. So for the ribbon worm tree the Yule model null hypothesis is rejected. For the uniform model null hypothesis test at the $\alpha = 0.05$ level the lower rejection limit is 5 cherries or less, and the upper rejection limit is 13 cherries or more, and so the test does not reject the uniform model null hypothesis. In any hypothesis test, however, it is important to note that a reconstructed tree is only an estimate of the underlying species tree. Consequently a more refined analysis would take into account the uncertainty and possible biases in phylogeny reconstruction [48, 41].

3.5 Some Extensions

Here we consider an extension of the EPU model for cherries. In particular we look at the asymptotic probability distribution for *triplets*, where a triplet is a cherry and a pendant edge adjacent to a common node. We look at both the rooted Yule model (Section 3.5.1) and the unrooted uniform model (Section 3.5.2).

In the construction of the generating matrix that follows we allow for the colour of the balls that are added back to follow a probability distribution, in which case the number in the generating matrix is the expected number of balls of that colour that are added back [66]. We also allow the generating matrix to be nonhomogeneous, meaning that the matrix entries may depend on the number of draws done, but with a matrix of constants as the limit as the number of draws became large [10].

3.5.1 Triplets For the Rooted Yule Model

Following the colouring scheme used in the urn model for cherries; let pendant edges that are part of cherries (but not triplets) be represented by black balls, pendant edges that are not part of triplets or cherries by white balls, and pendant edges that are part of triplets by green balls. The number of triplets is then one third of the number of green balls, and the total number of all coloured balls is equal to the number of leaves on the tree.

The following urn scheme generates the appropriate probability distribution for the

number of triplet edges. Starting with a rooted binary tree, put two black balls in the urn. Select a ball at random from the urn then return it. If the ball selected was white, take out a white ball and put in two black balls. If the ball selected was black then take out two black balls and put in three green balls. If the ball selected was green then there are two possibilities, depending on whether the ball corresponds to a pendant edge that is part of a cherry or not. One third of the time four black balls should be put in and three green balls removed. Two thirds of the time a white balls should be put in. So, on average, if the ball selected was green then 2/3 white balls should be put in, 4/3 black balls put in, and one green ball removed. Repeat the process of random selection and addition of balls $n-2$ time until there are n balls in the urn.

The generating matrix for this scheme is

$$A = \begin{pmatrix} -1 & 2 & 0 \\ 0 & -2 & 3 \\ 2/3 & 4/3 & -1 \end{pmatrix}, \tag{3.3}$$

where balls of type one are white, those of type two are black, and those of type three are green.

The eigenvalues of A are $-3, -2$ and the principal eigenvalue $\lambda_1 = 1$. For the principal eigenvalue the left eigenvector of A for which the entries add up to one is $v = (1/6\ 1/3\ 1/2)^T$. Let T_n be a random variable representing the number of triplets, then applying Theorem 5 we obtain the following result.

Corollary 4 *For the Yule model on rooted trees*

$$\frac{T_n - n/6}{\sqrt{k_1 n}} \to \mathcal{N}(0,1)$$

where k_1 is a constant.

A similar analysis for the balls coloured white and black reveals an interesting symmetry. The asymptotic probability distribution for the number of pendant edges that are not part of a cherry or triplet is normal with mean $n/6$, as is the asymptotic probability distribution for the number of cherries that are not part of triplets. It is expected that the variances differ, though in all cases the variance is, asymptotically, proportional to n.

3.5.2 Triplets for the Unrooted Uniform Model

As when we were dealing with cherries, the appropriate urn scheme for triplets in the uniform model is more complex because of the need to account for the internal edges. As will be shown, the relevant distinction that needs to made is between internal edges that are adjacent to cherries, and those that are not. Consequently, the following lemma will be useful in determining the generating matrix for triplets under the uniform model.

Lemma 5 *Let $\mathbb{P}_{int}(n)$ be the probability that a randomly selected internal edge from a randomly generated tree on n leaves is adjacent to a cherry. Under the uniform model on unrooted trees we have*

$$\mathbb{P}_{int}(n) = \frac{n(n-1)}{2(2n-5)(n-3)} .$$

Asymptotically, we have $\mathbb{P}_{int}(n) = 1/4$.

Proof. For a randomly generated tree with C_n cherries the proportion of internal edges adjacent to a cherry is, for $n \geq 5$,

$$\frac{C_n}{n-3} . \tag{3.4}$$

We have from from Theorem 7 that the expected number of cherries for a tree on n leaves is, under the uniform model,

$$\mathbb{E}[C_n] = \frac{n(n-1)}{2(2n-5)} .$$

Using this expected value in (3.4) gives $\mathbb{P}_{int}(n)$, and from this the asymptotic result follows directly. □

In the construction of the generating matrix that follows we allow for the colour of the balls that are added back to follow a probability distribution, in which case the number in the generating matrix is the expected number of balls of that colour that are added back [66]. We also allow the generating matrix to be nonhomogeneous, meaning that the entries depend on the number of draws done, but with a constant matrix as the limit as the number of draws became large. The relevant theory for nonhomogeneous generating

matrices was developed in Bai [10], where it was shown that the asymptotic probability distribution for the number of balls of each colour is normal.

Following the same colouring as was used in the previous section, let pendant edges that are part of cherries (but not triplets) be represented by black balls, pendant edges that are not part of triplets or cherries by white balls, and edges that are part of triplets by green balls. Furthermore, let internal edges be represented by red balls. The number of triplets is then one third of the number of green balls, and the total number of all coloured balls is equal to the number of *edges* of the tree.

Starting with an unrooted binary tree with four leaves, put four black balls and one red ball in an urn. Select a ball at random, note the colour, then return it. When the ball drawn is white, black, or green the same scheme for returning balls is followed as for the Yule model, except that a red ball is also added to the urn in each case (representing the formation of another internal edge). Thus the section of the generating matrix for the white, black, and green balls is the same as equation (3.3), but with a column of ones added to account for the red balls.

If the ball selected was red then there are two possibilities, depending on whether or not the internal edge the red balls represents had a cherry connected to it or not. If there was a cherry edge attached then two black balls should be removed, three green balls added, and a red ball added; otherwise a white ball and a red ball should be added. Using Lemma 5 then the fourth row of the generating matrix is, based on expected values,

$$\left(1-\frac{n(n-1)}{2(2n-5)(n-3)}\right)[\,1\;\;0\;\;0\;\;1\,]+\frac{n(n-1)}{2(2n-5)(n-3)}[\,0\;\;\text{-}2\;\;3\;\;1\,]$$

which equals

$$[1-\frac{n(n-1)}{2(2n-5)(n-3)}\quad\frac{-2n(n-1)}{2(2n-5)(n-3)}\quad\frac{3n(n-1)}{2(2n-5)(n-3)}\quad 1\;]\,.$$

Or asymptotically, as n becomes large,

$$[\,3/4\quad -1/2\quad 3/4\quad 1\,]\,.$$

So, asymptotically, the generating matrix is

$$A = \begin{pmatrix} -1 & 2 & 0 & 1 \\ 0 & -2 & 3 & 1 \\ 2/3 & 4/3 & -1 & 1 \\ 3/4 & -1/2 & 3/4 & 1 \end{pmatrix}.$$

However, because some of the entries are negative, this asymptotic generating matrix is of a form which does not satisfy the conditions which allow for an asymptotic analysis of the triplets distribution [10]. Even if we ignore that the fact that the matrix is the asymptotic form of a nonhomogenous matrix we run into difficulties since one of the negative elements is off the diagonal [66], a difficulty that cannot be removed by a relabelling of the colours. The best we can do is offer a conjecture, the conjecture being based on the speculation that theory similar to that in [10, 66] will be developed in the future. The pertinent feature of the asymptotic generating matrix A is that eventually the number of balls of each colour should go to infinity so similar results should apply (Z. Bai, personal communication). Calculating eigenvalues and eigenvectors we have: the principal eigenvalue of A is $\lambda_1 = 2$, with left eigenvector, for which the entries add up to one of $v = (\ 7/40\ \ 1/10\ \ 9/40\ \ 1/2\ )^T$. Following [10, 66] we have the following conjecture.

Conjecture 1 *For the uniform model on unrooted trees*

$$\frac{T_n - 3n/20}{\sqrt{k_2 n}} \to \mathcal{N}(0,1)$$

where k_2 is a constant.

Chapter 4

Rooting an Unrooted Tree

4.1 Introduction

Typically, construction of an evolutionary tree for a set of species is a two stage process. In the first stage, using biological data of some sort, an unrooted tree is constructed. In the next stage the unrooted tree is rooted along some edge. Commonly this is done by outgroup comparison, but embryological or fossil data can be used as well [57].

However, in some circumstances an outgroup is not available, or the embryological or fossil data is unclear. Furthermore, the choice of outgroup can strongly influence the accuracy of tree reconstruction [65]. In these circumstances heuristic methods have to be resorted to in order to root a tree. For example, in the *midpoint method*, the root is located at the point halfway between the two leaves that are the furtherest distance apart [23, 72]. In another approach the root is located at a point where the mean distance to the species on either side is the same (for example, the program TREECON [54] uses this method).

Here we present an approach for rooting a tree based on the shape of the tree and a simple probabilistic model for the growth of rooted trees (the Yule model). We show that even for large unrooted trees the approximate location of the edge that contains the root can be narrowed down to a small subset of edges.

4.2 Maximum Likelihood Method

If a probabilistic model for the growth of evolutionary trees is assumed then the edge(s) which contains the root for a tree can be found by using the *method of maximum likelihood.* The probabilistic model we consider here is the Yule model, chosen for its simplicity, but the same general approach can be used for more sophisticated models.

Let e be an edge of an unrooted labelled binary tree T. The conditional probability $\mathbb{P}[e \mid T]$ is the probability that the edge e contains the true root of T. Given an unrooted binary tree T, the method of maximum likelihood selects as its estimate of the root edge any edge e that maximises $\mathbb{P}[e \mid T]$. We let $E_{\max}(T)$ denote the set of edges of T that maximise $\mathbb{P}[e \mid T]$, and we let $e_{\max}(T)$ denote any edge in $E_{\max}(T)$. It is possible, for example when symmetry is present, that $|E_{\max}(T)| > 1$, but we will show below that, for the Yule model, $|E_{\max}(T)| \leq 3$.

For computational purposes $\mathbb{P}[e \mid T]$ may be expanded so that

$$\mathbb{P}[e \mid T] = \frac{\mathbb{P}[e, T]}{\mathbb{P}[T]}, \tag{4.1}$$

where $\mathbb{P}[e, T]$ is the probability of the labelled rooted tree obtained by rooting T on the edge e. For example, consider the labelled unrooted tree on 4 leaves (Figure 4.1). The probability of this tree ($\mathbb{P}[T]$) is 1/3. For the interior edge, the probability of the corresponding labelled rooted tree is 1/9, thus the conditional probability for the interior edge is 1/3. For the pendant edges the probability of the corresponding labelled rooted trees are all 1/18, thus the conditional probability for each pendant edge is 1/6.

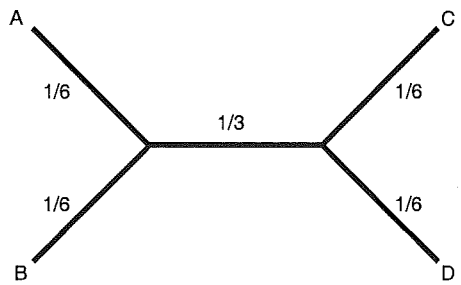


Figure 4.1: Conditional probabilities ($\mathbb{P}[e \mid T]$) for the edges of a labelled unrooted tree on 4 leaves.

Before we investigate the maximum likelihood probabilities for larger trees we state

some lemmata regarding what edges can be in $E_{\max}(T)$, and the size of this set.

Lemma 6 *Let edge e be an internal edge of an unrooted binary phylogenetic tree T. Denote the four subtrees of T adjacent to e by A, B, C, D, and let a, b, c, d respectively denote the number of leaves in these trees (Figure 4.2b). Let H(e) be the number of possible histories for the tree rooted along edge e. Then H(e) ≥ H(e′) for each of the four edges e′ incident with e precisely if both the following two inequalities hold:*

$$a + b \geq max\{c, d\}; \quad c + d \geq max\{a, b\}. \tag{4.2}$$

Furthermore, H(e) > H(e′) for all e′ precisely if these two inequalities hold as strict inequalities.

Proof. Without loss of generality we may represent e and e' as in Figure 4.2a. From (1.21) for the number of histories we have

$$H(e) = \frac{(n-1)!}{(n-1)(c+d-1)\prod_{v\in\mathring{C}}\delta(v)\prod_{v\in\mathring{D}}\delta(v)\prod_{v\in\mathring{F}}\delta(v)}. \tag{4.3}$$

If the tree is rooted at the adjacent edge e' the number of histories is

$$H(e') = \frac{(n-1)!}{(n-1)(f+d-1)\prod_{v\in\mathring{C}}\delta(v)\prod_{v\in\mathring{D}}\delta(v)\prod_{v\in\mathring{F}}\delta(v)}. \tag{4.4}$$

Therefore, $H(e) \geq H(e')$ precisely if

$$f \geq c. \tag{4.5}$$

Now let the tree F be split into two subtrees A and B (Figure 4.2b). Applying (4.5) to the edge labelled e, and then labeling in turn each adjacent edge as e', leads to the two 4-branch inequalities. If both of the 4-branch inequalities are strict then $H(e)$ is strictly larger than $H(e')$. □

Lemma 7 *Any two edges in $E_{\max}(T)$ are adjacent.*

Proof. We will derive a contradiction by supposing that there exists two non-adjacent edges e_1, e_2 in $E_{\max}(T)$. Under this assumption we can represent T as in Figure 4.2c,

where $k \geq 1$, and $c_0, c_1, \ldots, c_k$ are all positive. For edge e_1 to be in $E_{\max}(T)$ we must have, from Lemma 6,

$$a + b \geq c + d + c_1 + \ldots + c_k. \tag{4.6}$$

Likewise for edge e_2 we must have

$$c + d \geq a + b + c_0 + \ldots + c_{k-1}. \tag{4.7}$$

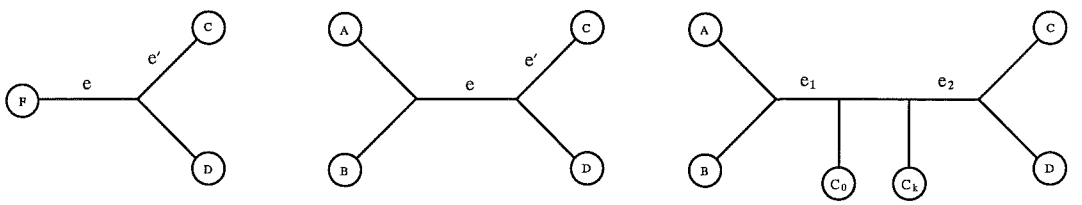
Adding (4.6) and (4.7) we get

$$a + b + c + d \geq a + b + c + d + c_0 + 2(c_1 + \ldots c_{k-1}) + c_k. \tag{4.8}$$

This implies that $c_0 = c_1 = \ldots = c_k = 0$, contradicting our original supposition, thus any two edges in $E_{\max}(T)$ must be adjacent. □

As we are dealing with binary trees we have the following straightforward consequence of Lemma 7.

Corollary 5 $|E_{\max}(T)| \leq 3$. *Furthermore, if both the inequalities in* (4.2) *are strict, then* $|E_{\max}(T)| = 1$.



(a) (b) (c)

Figure 4.2: Generic unrooted binary trees with subtrees A, B, C, D, F with a, b, c, d, f leaves respectively. (a) With three distinguished edges (b) With four distinguished edges (c) A hypothetical tree with two edges e_1, e_2 in $E_{\max}(T)$ that are separated by $k \geq 1$ edges. $C_0, \ldots, C_k$ denotes subtrees with $c_0, \ldots, c_k$ leaves, respectively.

4.3 Mean Probability of Finding the True Root

Given an unrooted labelled binary tree T let e_{max} be the edge for which $\mathbb{P}[e \mid T]$ is maximum. The probability that e_{max} contains the true root is $\mathbb{P}[e_{max} \mid T]$. If T is obtained by generating a rooted tree according to the Yule model, then unrooting the tree, let $\epsilon(n)$ denote the mean probability that e_{max} contains the true root. Then,

$$\epsilon(n) = \sum \mathbb{P}[e_{max} \mid T]\, \mathbb{P}[T]\,, \tag{4.9}$$

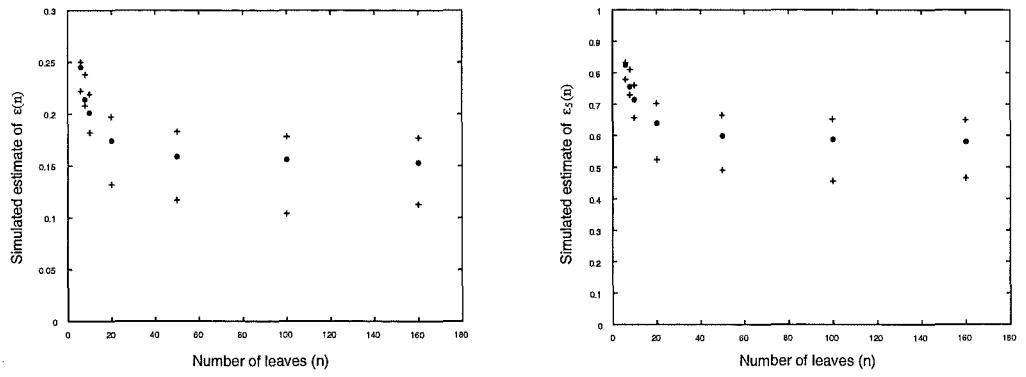
where is the summation is over all labelled unrooted trees T on n leaves. For n small $\epsilon(n)$ can be explicitly calculated, but for larger values $\epsilon(n)$ has to be approximated by simulation. Simulated values were calculated by the formula

$$\epsilon(n) \approx \frac{1}{N}\sum_{i=1}^{N} \mathbb{P}[e_{max} \mid T_i] = \frac{1}{N}\sum_{i=1}^{N} \frac{\mathbb{P}[e_{max}, T_i]}{\mathbb{P}[T_i]}\,, \tag{4.10}$$

where T_i is a labelled unrooted binary tree on n leaves obtained by generating a rooted tree according to the Yule process, then unrooting it. N is the number of trees generated in the simulation.

The simulation results suggest that $\lim_{n\to\infty} \epsilon(n) \approx 0.15$ (Figure 4.3a). The five edges with the largest conditional probabilities for a tree are always an interior edge and the four edges adjacent to it. Let $\epsilon_5(n)$ denote the mean value for the sum of the five largest conditional probabilities for a tree. The simulations suggest that $\lim_{n\to\infty} \epsilon_5(n) \approx 0.58$ (Figure 4.3b). Thus, even for a large unrooted tree, the location of the root may be narrowed down to a small cluster of five edges, of which one is more likely than not to be the true root. That the asymptotic conditional probability is nonzero concurs with the behaviour for an analogous model, the Yule-Furry model, in which edges are added at random to nodes [34].

The asymptotic mean probability can be found by embedding the discrete process of rooting a tree into a continuous analogue involving 'stick breaking'. The asymptotic properties of this process have been previously analysed in [78]; here we are concerned only with comparisons involving the first two breaks of the stick. For the proof of the following theorem see Steel and McKenzie in [68].



(a) (b)

Figure 4.3: Simulation results for the conditional probability of edges. Two hundred unrooted trees were randomly generated for different values of n. The trees were produced by unrooting the rooted tree generated by a Yule process. The minimum and maximum probabilities for each simulation are represented by crosses (+), and the mean by a dot (•). (a) Estimate of the mean probability that e_{max} contains the true root. (b) Estimate of the mean value for the sum of the five largest conditional probabilities for a tree.

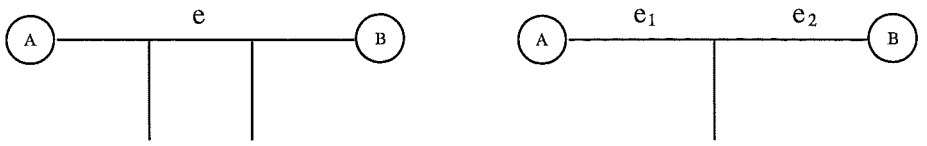
Theorem 8 *[68] The edge e_{max} chosen by the method of maximum likelihood has asymptotic mean conditional probability of* $4\ln(4/3) - 1 \approx 0.15$.

4.4 Lower Bound

From the simulation results it appears that the tree for which $\mathbb{P}[e_{max} \mid T]$ is smallest is the unrooted caterpillar tree, which is the tree shape obtained from unrooting the rooted comb shape (see Section 1.5.4). For the unrooted caterpillar tree $\mathbb{P}[e_{max} \mid T]$ may be calculated exactly, and furthermore it can be shown that asymptotically this probability is zero. Before we do this we must determine what edge(s) are in $E_{\max}$ for the caterpillar tree. The following lemma gives the required result.

Lemma 8 *Let CT_n be the unrooted caterpillar tree on n leaves.*

(i) For n even there is a single edge in $E_{\max}(CT_n)$, and it is located at the internal edge where there are $\frac{n}{2}$ leaves on each side.



(a) (b)

Figure 4.4: Generic unrooted caterpillars. A and B are subtrees with a, b leaves respectively, where a and b sum to an even number. (a) Even number of leaves. (b) Odd number of leaves.

(ii) For n odd there are two edges in $E_{\max}(CT_n)$, located at the two internal edges where there are $\frac{n-1}{2}$ leaves on one side and $\frac{n+1}{2}$ leaves on the other side.

Proof. For n even, consider the generic caterpillar where the subtrees A, B have a, b leaves respectively (Figure 4.4a). Applying the 4-branch conditions from Lemma 6 to the edge labelled e gives

$$b \leq a + 1$$
$$b \geq a - 1 .$$

Since a and b are integers, and their sum is even, we must have $a = b = \frac{n}{2} - 1$.

For n odd, consider the generic caterpillar tree where the subtrees A, B have a, b leaves respectively (Figure 4.4b). Applying the 4-branch conditions to the edge labelled e_1 gives

$$b \leq a$$
$$b \geq a - 1.$$

Since a and b are integers, and their sum is even, we must have $a = b = \frac{n-1}{2}$. By symmetry, applying the 4-branch conditions to the edge e_2 gives the same constraints on a and b. Thus both e_1 and e_2 are in $E_{\max}$ for n odd. □

Theorem 9 *The edge(s) (e_{max}) chosen by the method of maximum likelihood for the unrooted labelled caterpillar tree on n leaves (CT_n) has conditional probability:*

$$\mathbb{P}[e_{max} \mid CT_n] = \begin{cases} \frac{8}{3}\frac{(n-2)!}{2^n(\frac{n-1}{2})!(\frac{n-3}{2})!}\,, & n \text{ odd} \\ \\ \frac{8}{3}\frac{(n-2)!}{2^n\{(\frac{n-2}{2})!\}^2}\,, & n \text{ even.} \end{cases}$$

Asymptotically, as $n \to \infty$,

$$\mathbb{P}\left[e_{max} \mid CT_n\right] \sim \frac{2}{3}\sqrt{\frac{2}{\pi}}\frac{1}{\sqrt{n}} \to 0\,.$$

Proof. Let CT_n be the labelled unrooted caterpillar tree on n leaves, with probability $\mathbb{P}[CT_n]$. Also let e_{max} be the edge(s) for which $\mathbb{P}[e \mid CT_n]$ is maximum. For the tree CT_n the location of the e_{max} edge(s) may easily be found (see Lemma 8). For n odd there are two e_{max} edges, located at the two edges where there are $\frac{n-1}{2}$ leaves on one side and $\frac{n+1}{2}$ leaves on the other side. For n even there is a single e_{max} edge, located at the edge where there is $\frac{n}{2}$ leaves on both sides. The probability that e_{max} contains the true root of CT_n is, in either case,

$$\mathbb{P}\left[e_{max} \mid CT_n\right] = \frac{\mathbb{P}[e_{max}, CT_n]}{\mathbb{P}[CT_n]}\,. \tag{4.11}$$

Consider, firstly, the numerator of this equation. From the explicit formula of equation (1.19), involving products over nodes, it follows that

$$\mathbb{P}\left[e_{max}, CT_n\right] = \begin{cases} \frac{2^{n-1}}{n!}\frac{1}{(n-1)(\frac{n-1}{2})!(\frac{n-3}{2})!}, & n \text{ odd} \\ \\ \frac{2^{n-1}}{n!}\frac{1}{(n-1)\{(\frac{n-2}{2})!\}^2}\,, & n \text{ even.} \end{cases}$$

Now consider the denominator of (4.11). The unlabelled unrooted caterpillar on n leaves (UCT_n) can only be obtained by adding edges to the four 'end' pendant edges of the unlabeled caterpillar on $n-1$ leaves. Hence the unlabeled caterpillar probabilities satisfy the recursion

$$\mathbb{P}[UCT_n] = \frac{4}{n-1}\mathbb{P}[UCT_{n-1}], \qquad \text{where } \mathbb{P}[UCT_5] = 1.$$

Solving this recursion, then dividing by the number of labelled caterpillars on n leaves ($\frac{n!}{8}$), gives

$$\mathbb{P}[CT_n] = \frac{3}{16}\frac{4^{n-2}}{n!(n-1)!}.$$

Combining the numerator and denominator terms gives the first part of the theorem. The second part of the theorem follows from the asymptotic equation $\frac{1}{2^n}\binom{n}{n/2} \sim \sqrt{2/(\pi n)}$.

□

Chapter 5

The Enumeration of Compatible Rooted Trees

5.1 Introduction

By $[n]$ we mean $\{1, 2, \ldots, n\}$. Let $S = \{T_1, T_2, \ldots, T_k\}$ be a set of labelled trees with leaf sets $L = \{\mathcal{L}_1, \mathcal{L}_2, \ldots, \mathcal{L}_k\}$ respectively such that each $\mathcal{L}_i$ is a proper subset of $[n]$. The set S is said to be *consistent* if there exists a phylogenetic tree T such that $T_{|\mathcal{L}_1} = T_1$, $T_{|\mathcal{L}_2} = T_2$, ..., $T_{|\mathcal{L}_k} = T_k$, in which case the tree T is said to be *compatible* with the set of trees S. If the trees in S are unrooted then it has been shown that the problem of determining whether or not there exists a compatible tree is NP-complete [11, 67], which means that the problem belongs to a set of problems for which it is suspected that there are no polynomial algorithms for their solution. If the trees in S are rooted then a solution to the problem of the existence of a compatible tree can be solved in polynomial time and efficient algorithms exist for the construction of the tree(s) [1, 18, 50].

Here we restrict ourselves to the simpler, but still illuminating case, where the set of trees in S are rooted, and the set L forms a partition of $[n]$. A set of rooted trees where the leaf sets form a partition is always consistent. For instance, the tree $T = T_1 + T_2 + \cdots + T_k$, where the root node has k edges connected to it, is a tree compatible with S in this case. Furthermore, even if T is restricted to being a binary tree, the set of trees are consistent. For example, one way of constructing a compatible tree that is binary is to start with the tree T_1 then attach on to any of its edges the tree T_2 (connecting the root of T_2 to T_1 with an edge). Continuing attaching the remaining trees

T_i in this way up to the tree T_k gives a compatible tree that is binary.

As is clear from this method of construction there is always more than one tree compatible with a set of trees S if the leaf sets form a partition. An example of a set of two rooted trees compatible with at least two trees is shown in Figure 5.1. A natural question to ask is just how many labelled rooted trees are there that are compatible with a set of labelled rooted trees with partitioned leaf sets? Previous work that dealt with unrooted trees suggests that there are no simple formulae for this enumeration problem [17]. We derive a recursive formula for the number of trees compatible with two leaf-set partitioned trees (Section 5.2). A program was written to evaluate the recursion and investigate some of its properties. Under the uniform model, we find the mean number of trees compatible with two leaf-set partitioned trees on r and s leaves. It was found that if both trees were caterpillar/symmetric then the number of compatible trees exceeds the mean number of compatible trees. We extend this work further, deriving a recursion for the number of labelled trees compatible with three leaf-set partitioned trees (Section 5.3). As will be shown, even for just three trees, a recursive formula for the number of compatible trees become unwieldy, and we show that the number of terms in such a recursion grows at the very least exponentially.

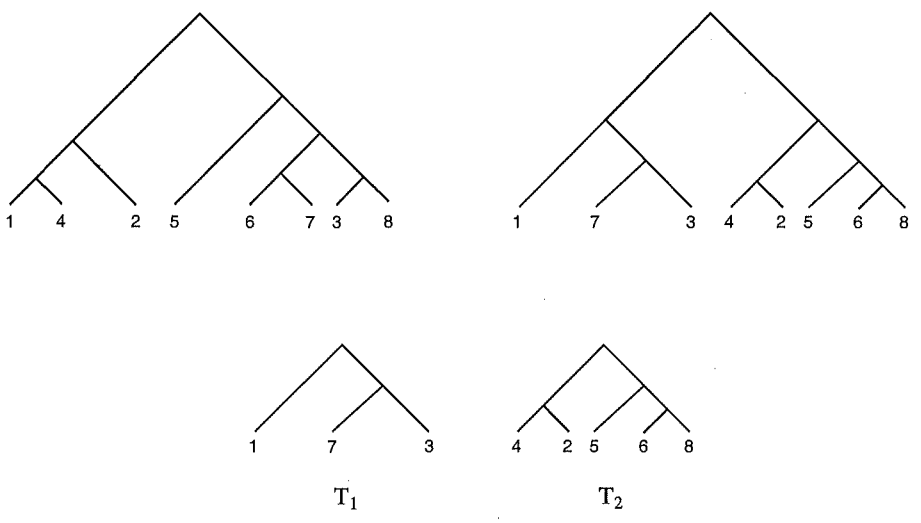


Figure 5.1: Two trees compatible with the tree T_1 and T_2.

5.2 Two Trees

5.2.1 A Recursion

For the case of two trees a recursive formula may be derived that gives the number of labelled trees compatible with both. Let the two trees be $T_1 = a_1 + b_1$ and $T_2 = a_2 + b_2$ (see Figure 5.2). We have the following proposition.

Proposition 1 *If $N(T_1, T_2)$ is the number of rooted trees compatible with T_1 and T_2 then*

$$\begin{aligned} N(T_1, T_2) = {} & N(a_1, a_2)N(b_1, b_2) + N(a_1, b_2)N(a_2, b_1) \\ & + N(a_1, T_2) + N(b_1, T_2) \\ & + N(a_2, T_1) + N(b_2, T_1) \\ & + 1 \,. \end{aligned} \tag{5.1}$$

Proof. The recursion formula summarises the seven different ways in which the subtrees of T_1 and T_2 may be joined together to form a compatible tree. Let the root vertex of a compatible tree be denoted by ρ_c. For the first term, consisting of a product, the subtrees in the sets $\{a_1, a_2\}$ and $\{b_1, b_2\}$ are on opposite sides of ρ_c. In the second term the subtrees in the sets $\{a_1, b_2\}$ and $\{a_2, b_1\}$ are on opposite sides of ρ_c. The third term arises from having b_1 on one side of ρ_c, while embedding a_1 in T_2 on the other side. The fourth term arises in a similar manner, but embedding b_1 in T_2 instead. The fifth and sixth terms come about in a similar manner to the third and fourth terms, but this time involving the subtrees a_2 and b_2. The final term arises by simply having T_1 on one side of ρ_c and T_2 on the other side. □

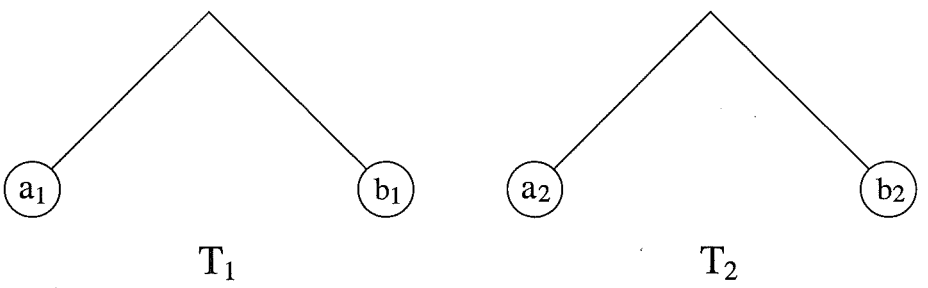


Figure 5.2: Tree compatibility for two trees. Tree T_1 has subtrees a_1 and b_1. Tree T_2 has subtrees a_2 and b_2.

This recursion was evaluated using a program written in *MATLAB* to give Table 5.1. As can be seen from the table the number of compatible trees grows rapidly as the number of leaves on the trees T_1 and T_2 increase. For example, if the trees T_1 and T_2 are the symmetric trees on eight leaves then the number of compatible trees exceeds one million.

Shape of T_1	Shape of T_2	$N(T_1, T_2)$
1	1	1
2	1	3
2	2	15
3	2	35
3	3	105
4_1	2	63
4_1	3	231
4_1	4_1	607
4_1	4_2	575
4_2	2	63
4_2	3	231
4_2	4_2	703
⋮	⋮	⋮
8_1	8_1	531 327
8_1	8_{12}	359 487
8_{12}	8_{12}	416 031
8_{12}	8_{23}	308 735
8_{23}	8_{23}	1 027 839

Table 5.1: Number of compatible trees for the labelled trees T_1 and T_2. The symbols for the tree shapes follow the 'dictionary' notation (see Section 1.3.2). In this notation a tree shape on n leaves has the symbol n_i, where $i = 1$ for the caterpillar shape, and i takes its largest value for the most symmetric tree shape.

5.2.2 Mean Value Under the Uniform Model

As a simple measure of the general trend in the number of compatible trees we calculate the mean value of $N(T_1, T_2)$ under the uniform model. We treat T_1 and T_2 as random variables, where T_1 has r leaves and T_2 has s leaves ($n = r + s$). Before we find the mean value we need the following lemma (for the unrooted case see [17]).

Lemma 9 *Let t be a labelled tree on k leaves with leaf set $\mathcal{L}$. The number of labelled trees T on n leaves such that $T|_{\mathcal{L}} = t$ is given by*

$$\frac{(2n-3)!!}{(2k-3)!!}. \tag{5.2}$$

Proof. Firstly, due to exchangeability, we can assume that t has the leaf set $\mathcal{L} = \{1, 2, \ldots, k\}$. We now add another $n - k$ leaf-labelled edges to form a labelled tree T on n leaves. The first edge, labelled $k + 1$ may be added in $2k - 1$ places (allowing for a 'ghost' edge at the root). The next edge, labelled $k + 2$, may be added in $2k + 1$ places. Continuing in this way, we can add $n - k$ leaf-labelled edges in $(2k - 1)(2k + 1)\ldots(2n - 3) = (2n - 3)!!/(2k - 3)!!$ ways, each of these representing a different labelled tree on n leaves. □

Let $\overline{N}_{r,s}$ be the expected value of $N(T_1, T_2)$ where T_1 and T_2 are randomly generated labelled trees on r and s leaves respectively ($n = r+s$). We have the following proposition.

Proposition 2 *For the uniform model*

$$\overline{N}_{r,s} = \frac{(2n-3)!!}{(2r-3)!!\,(2s-3)!!}. \tag{5.3}$$

Proof. We have, by the definition of expectation,

$$\overline{N}_{r,s} = \sum_{i,j} N(T_i, T_j)\mathbb{P}(T_i, T_j)$$

where the summation is over all labelled trees T_i on r leaves and labelled trees T_j on s leaves. $\mathbb{P}(T_i, T_j)$ is the probability of obtaining the particular trees T_i and T_j. We have, for the uniform model, $\mathbb{P}(T_i, T_j) = \frac{1}{(2r-3)!!\,(2s-3)!!}$ so

$$\overline{N}_{r,s} = \frac{1}{(2r-3)!!\,(2s-3)!!} \sum_{i,j} N(T_i, T_j) \; .$$

Breaking the summation into two parts gives

$$\begin{aligned} \overline{N}_{r,s} &= \frac{1}{(2r-3)!!\,(2s-3)!!} \sum_{i} \sum_{j} N(T_i, T_j) \\ &= \frac{1}{(2r-3)!!\,(2s-3)!!} \sum_{i} \frac{(2n-3)!!}{(2r-3)!!} \qquad \text{(Lemma 9)} \\ &= \frac{(2n-3)!!}{(2r-3)!!\,(2s-3)!!} \; . \end{aligned}$$

□

5.2.3 Maximising the Number of Compatible Trees

From Proposition 2, the values of r and s that maximise $\overline{N}_{r,s}$, for a given value of n (where $n = r + s$) can be found. For n even there is a single maxima at $r = s = n/2$. For n odd there are two maxima: $r = (n-1)/2,\ s = (n+1)/2$ and $r = (n+1)/2,\ s = (n-1)/2$. Based on this, and computer simulations using the recursive formula for the number of compatible trees (Proposition 1), we conjecture that, for a given value of n, the maximum number of compatible trees results when the trees have the minimum difference in the number of leaves and are symmetric. This, of itself, is not that surprising. What is more surprising, as we shall show, is the extent to which the maximum number of compatible trees exceeds the mean number of compatible trees. Furthermore, as we shall also show, trees that are of the same size and that are both caterpillars also have a higher than average number of compatible trees. The simulations also suggest that having one tree a caterpillar, and the other symmetric, is associated with a lower than average number of compatible trees.

In its most general form the recursion in Proposition 1 is quite intricate to program, and slow to run. However, if both T_1, T_2 are caterpillars or fully symmetric then the recursion simplifies considerably, resulting in a much simpler and faster program. For simulations with just fully symmetric trees or caterpillars we used the results in the following corollary.

Corollary 6 *If C_k, C_l are labelled caterpillar trees on k, l leaves respectively then we have*

$$N(C_k, C_l) = N(C_k, C_{l-1}) + N(C_{k-1}, C_{l-1}) + N(C_{k-1}, C_l) + 4(k-1)(l-1) + 4 .$$

If S_k, S_l are labelled fully symmetric trees on $2^k, 2^l$ leaves respectively then we have

$$N(S_k, S_l) = 2\left[N(S_k, S_{l-1}) + N(S_{k-1}, S_{l-1})^2 + N(S_{k-1}, S_l)\right] + 1 .$$

Proof. Starting with the caterpillar trees C_k, C_l then from Proposition 1 we have

$$\begin{aligned} N(T_k, T_l) = {} & N(1,1)N(T_{k-1}, T_{l-1}) + N(1, T_{l-1})N(T_{k-1}, 1) \\ & + N(1, T_l) + N(T_{k-1}, T_l) \\ & + N(1, T_k) + N(T_{l-1}, T_k) \\ & + 1 . \end{aligned}$$

Since for a tree on n leaves there is $2n-1$ places to add a labelled edge, each giving a different labelled tree, then we have

$$\begin{aligned} N(T_k, T_l) = {} & N(T_{k-1}, T_{l-1}) + [2(l-1) - 1] + [2(k-1) - 1] \\ & + (2l - 1) + N(T_{k-1}, T_l) \\ & + (2k - 1) + N(T_{l-1}, T_k) \\ & + 1 , \end{aligned}$$

and simplifying this gives the required result for caterpillars. If we start with the fully symmetric trees S_1, S_2 then using Proposition 1 again we have

$$\begin{aligned} N(S_k, S_l) = {} & N(S_{k-1}, S_{l-1})N(S_{k-1}, S_{l-1}) + N(S_{k-1}, S_{l-1})N(S_{k-1}, S_{l-1}) \\ & + N(S_{k-1}, S_l) + N(S_{k-1}, S_l) \\ & + N(S_{l-1}, S_l) + N(S_{l-1}, S_k) \\ & + 1 . \end{aligned}$$

Simplifying this give the recursion for fully symmetric trees. □

Using Corollary 6 we calculated the ratio of the number of compatible trees for two caterpillar/fully symmetric trees to the mean number of compatible trees for two trees with the same number of leaves. The results are shown in Table 5.2. The ratios in both cases seem to be going off to infinity, though at a slower rate for the caterpillar trees.

n	Fully symmetric trees	Caterpillar trees
2	1.0	1.0
4	1.2	1.0
8	3.0	1.6
16	45.4	10.8
32	21876.8	1553.4
64	1.05×10^{10}	9.28×10^{7}

Table 5.2: For two trees on n leaves, the ratio of the number of compatible trees if they were both fully symmetric/caterpillar trees to the mean number of compatible trees for two trees on n leaves.

5.3 Three Trees

Continuing in the same vein as the previous section, we derive a recursion for the number of rooted trees compatible with three trees for which the leaf-sets form a partition of $[n]$. As will be seen, the recursion for three trees is considerably more complicated than the recursion for two trees. Let the three trees be $T_1 = a_1 + b_1$, $T_2 = a_2 + b_2$, $T_3 = a_3 + b_3$. Also let $N(T_1, T_2, T_3)$ be the number of trees compatible with T_1, T_2, T_3. We have the following proposition.

Proposition 3

$$N(T_1, T_2, T_3) = F_{10} + F_{20} + F_{22} + F_{30} + F_{32} \ ,$$

where F_{ij} $(1 \leq i \leq 3,\ j = 0, 2 : j \leq i)$ *denotes a collection of terms for which the compatible trees they are counting has a left subtree containing* i *of the subtrees* $\{a_1, a_2, a_3, b_1, b_2, b_3\}$, *of which* j *of these are from the same tree (one of* T_1, T_2, T_3*). The right subtree of the compatible tree being counted contains the remaining* $6 - i$ *subtrees. The* F_{ij} *are:*

$$
\begin{aligned}
F_{10} &= N(b_1, T_2, T_3) + N(a_1, T_2, T_3) + N(b_2, T_1, T_3) \\
&\quad + N(a_2, T_1, T_3) + N(b_3, T_1, T_2) + N(a_3, T_1, T_2) \,,
\end{aligned} \tag{5.4}
$$

$$
\begin{aligned}
F_{20} &= N(a_1, a_2)N(b_1, b_2, T_3) + N(a_1, b_2)N(b_1, a_2, T_3) \\
&\quad + N(a_1, a_3)N(b_1, b_3, T_2) + N(a_1, b_3)N(b_1, a_3, T_2) \\
&\quad + N(b_1, a_2)N(a_1, b_2, T_3) + N(b_1, b_2)N(a_1, a_2, T_3) \\
&\quad + N(b_1, a_3)N(a_1, b_3, T_2) + N(b_1, b_3)N(a_1, a_3, T_2) \\
&\quad + N(a_2, a_3)N(b_2, b_3, T_1) + N(a_2, b_3)N(b_2, a_3, T_1) \\
&\quad + N(b_2, a_3)N(a_2, b_3, T_1) + N(b_2, b_3)N(a_2, a_3, T_1) \,,
\end{aligned} \tag{5.5}
$$

$$
F_{22} = N(a_1, b_1)N(T_2, T_3) + N(a_2, b_2)N(T_1, T_3) + N(a_3, b_3)N(T_1, T_2) \,, \tag{5.6}
$$

$$
\begin{aligned}
F_{30} &= N(a_1, a_2, a_3)N(b_1, b_2, b_3) + N(a_1, a_2, b_3)N(b_1, b_2, a_3) \\
&\quad + N(a_1, b_2, a_3)N(b_1, a_2, b_3) + N(b_1, a_2, a_3)N(a_1, b_1, b_3) \,,
\end{aligned} \tag{5.7}
$$

$$
\begin{aligned}
F_{32} &= N(T_1, a_2)N(b_2, T_3) + N(T_1, b_2)N(a_2, T_3) \\
&\quad + N(T_2, a_1)N(b_1, T_3) + N(T_2, b_1)N(a_1, T_3) \\
&\quad + N(T_2, a_1)N(b_1, T_3) + N(T_3, b_1)N(T_2, a_2) \,.
\end{aligned} \tag{5.8}
$$

Proof. The recursion summarises the 31 different ways that the subtrees $a_1, a_2, a_3, b_1, b_2, b_3$ may be joined together to form a tree compatible with T_1, T_2, T_3. Let ρ_c be the root node of such a compatible tree. For the first term F_{10} we have just a single subtree on one side of the root, while all the rest are on the other side. In F_{20} we have two subtrees on one side of ρ_c, of which neither is from the same tree (one of T_1, T_2, T_3), while all the other subtrees are on the other side of the root. The term F_{22} is similar to F_{20}, but the two subtrees on one side of ρ_c are from the same tree. The term F_{30} accounts for the trees in which there is three subtrees on one side of ρ_c, of which no pair are from the same tree, while all other subtrees are on the other side of ρ_c. The term F_{32} is similar to F_{30}, but accounts for the trees in which of the three subtrees on one side there is a pair of them from the same tree. □

5.4 Discussion

For two trees a recursion with 7 terms was required, while for three trees the recursion had 31 terms in it. Thus, as is apparent from these two cases, the recursions for more than three trees are going to involve a large number of terms. This can be quantified, which we do here by calculating a lower bound for the number of terms. For a collection of more than three trees the recursion for the number of compatible trees will involve terms of the form F_{ij} already given, and other terms accounting for additional combinations of subtrees. So if we count the number of terms in $\{F_{10}, F_{20}, F_{22}, F_{30}, F_{32}\}$, where we are dealing with a collection of m trees, this will give a lower bound for the total number of terms required. Crucially we also count the number of terms in F_{m0}, which is the collection of terms for which there are m subtrees from $\{a_1, b_1, \ldots, a_m, b_m\}$ on one side of the root, of which none are from the same tree.

Proposition 4 *Let $\mathcal{NT}(m)$ be the number of terms required to recursively calculate the number of compatible trees for a collection of m rooted trees whose leaf sets form a partition of $[n]$. We have, for $m \geq 4$,*

$$\mathcal{NT}(m) > 2^{m-1} + 8m^3 - 20m^2 + 15m\,. \tag{5.9}$$

Proof. Clearly the number of terms associated with F_{10} is $2m$, and for F_{22} is m. For the term F_{20} we have two subtrees not from the same tree on one side, with the rest of the subtrees on the other side. From $2m$ subtrees we can select $\binom{2m}{2}$ combinations of two subtrees at a time, of which m will be combinations in which the subtrees are from the same tree, thus we have $\binom{2m}{2} - m$ terms associated with F_{20}.

For the term F_{30} we must have three subtrees of which none are from the same tree. We can select the first subtree in $2m$ ways, the second in $2m-2$ ways, and the third in $2m-4$ ways, giving a total of $8m(m-1)(m-2)$ ways. For the term F_{32} we have m ways of picking two subtrees so that they are from the same tree, and $2m-2$ ways of selecting another subtree to go with the first two, thus giving $2m(m-1)$ terms associated with F_{32}. Note that for the special case of $m = 3$ we have $m(m-1)$ terms because of symmetry that is not present for higher values of m.

Finally, for the term F_{m0}, for each subtree a_i on one side of the root we must have the subtree b_i on the other side. For each pair there is two ways that this may occur, so for

m pairs there is, if we disregard symmetry about the root, 2^m ways. Taking into account symmetry about the root gives 2^{m-1} terms associated with F_{m0}.

Adding up all the terms associated with each F_{ij} we have mentioned then we obtain the above lower bound for $\mathcal{NT}(m)$. □

Calculating this lower bound then we get that, for example, $\mathcal{NT}(4) > 260$ and $\mathcal{NT}(5) > 591$, so the number of terms required in the recursions for trees on even a small number of leaves is large. However, the main point to be obtained from the lower bound is that $\mathcal{NT}(m) > 2^{m-1}$, so the number of terms required to recursively calculate the number of compatible trees for a collection of m leaf-set partitioned trees grows at the very least exponentially.

Chapter 6

Maximum Agreement Subtrees (MASTs)

6.1 Introduction

Let T and T' be labelled rooted trees on n leaves. Also let $S \subseteq [n] : |S| = s$. Then t is an agreement subtree of T and T' if $T_{|S} = T'_{|S} = t$ for some S. If t is an agreement subtree for which the value of s is maximum then it is a maximum agreement subtree (MAST).

MASTs offer a way of summarising what information a set of trees have in common and were introduced by Finden and Gordon [25]. Calculating a MAST for two trees can be achieved by a polynomial time algorithm [70], but has been shown to be a NP-hard problem for a set of three or more trees [5]. In this chapter we are not concerned with algorithms for finding a MAST, but instead our interest is in the *size* (number of leaves) of a MAST for two trees. In previous work on this problem lower bounds on the size of a MAST for two trees have been found for some special cases, the depth (maximum root node to leaf distance) of each tree being an important factor [29]. In this chapter, taking a different approach, we investigate the probability that the size of a MAST for two randomly generated binary trees exceeds a certain value.

6.2 Upper Bound

Here we derive an upper bound for the probability that two randomly generated trees have a MAST of size greater than or equal to s. In the following two sections we determine

this upper bound explicitly for the uniform model and recursively for the Yule model.

For a leaf subset S and trees T and T' a useful indicator variable is X_S defined by

$$X_S = \begin{cases} 1, & \text{if } T_{|S} = T'_{|S} \\ 0, & \text{otherwise .} \end{cases} \tag{6.1}$$

The number of agreement subtrees with s leaves for T and T' is then counted by

$$X^{(s)} = \sum_{S \subseteq [n]:|S|=s} X_S \, . \tag{6.2}$$

A desirable quantity to know is the maximum value that s may take, this being the number of leaves on a MAST. Or, if the trees T and T' are randomly generated, what is the probability that a MAST has a size exceeding a certain value. That is,

$$\mathbb{P}[T \text{ and } T' \text{ have a MAST of size } \geq s] \, . \tag{6.3}$$

If the trees T and T' are randomly generated then X_S defines a class of random indicator variables with $\binom{n}{s}$ members, each member indexed by the subscript S. Because tree probabilities are invariant under leaf relabelling, as are induced subtree relationships, then all members of the class defined by X_S have the same probability distribution. Similarly, for randomly generated trees, $X^{(s)}$ defines a class of random counting variables with n members, indexed by the parameter s, but the members of the class have different probability distributions. We now give an upper bound for (6.3).

Theorem 10 *Let T and T' be two randomly generated labelled rooted trees on n leaves. We have*

$$\mathbb{P}[T \textit{ and } T' \textit{ have a MAST of size } \geq s] \leq \binom{n}{s} \sum_{t \in RB(s)} \mathbb{P}_s[t]^2 \, ,$$

where $RB(s)$ is the set of labelled rooted trees on s leaves.

Proof. The event $\{T \text{ and } T' \text{ have a MAST of size } \geq s\}$ is equivalent to the event $\{X^{(s)} \geq 1\}$ so

$$\begin{aligned}
\mathbb{P}[T \text{ and } T' \text{ have a MAST of size } \geq s] &= \mathbb{P}[X^{(s)} \geq 1] \\
&\leq \mathbb{E}[X^{(s)}] && \text{(Markov inequality)} \\
&= \mathbb{E}[\sum_{S \subseteq [n]:|S|=s} X_S] && \text{(by definition)} \\
&= \sum_{S \subseteq [n]:|S|=s} \mathbb{E}[X_S] && \text{(property of expectation)} \\
&= \sum_{S \subseteq [n]:|S|=s} \mathbb{P}[X_S = 1] && \text{(indicator random variable)} \\
&= \binom{n}{s} \mathbb{P}[X_{\{1,2,\ldots,s\}} = 1] \,. && \text{(labelling invariance)}
\end{aligned}$$

Now we have,

$$\begin{aligned}
\mathbb{P}[X_{\{1,2,\ldots,s\}} = 1] &= \mathbb{P}_n[T_{|\{1,2,\ldots,s\}} = T'_{|\{1,2,\ldots,s\}}] && \text{(by definition)} \\
&= \sum_{t \in RB(s)} \mathbb{P}_n[T_{|\{1,2,\ldots,s\}} = t \text{ and } T'_{|\{1,2,\ldots,s\}} = t] && \text{(all possibilities)} \\
&= \sum_{t \in RB(s)} \mathbb{P}_n[T_{|\{1,2,\ldots,s\}} = t] \; \mathbb{P}_n[T'_{|\{1,2,\ldots,s\}} = t] && \text{(independence of } T \text{ and } T') \\
&= \sum_{t \in RB(s)} \mathbb{P}_n[T_{|\{1,2,\ldots,s\}} = t]^2 && (T \text{ and } T' \text{ have the same distribution}) \\
&= \sum_{t \in RB(s)} \mathbb{P}_s[t]^2 \,. && \text{(sampling consistency property)}
\end{aligned}$$

So, upon substituting back for this term, we obtain the upper bound as stated in the theorem. □

6.2.1 Uniform Model

We now find an analytical form for the upper bound under the uniform model.

Corollary 7 *For the uniform model on rooted trees we have*

$$\mathbb{P}[T \text{ and } T' \text{ have a MAST of size } \geq s] \leq \binom{n}{s} \frac{1}{(2s-3)!!} \,.$$

Proof.

$$\sum_{t \in RB(s)} \mathbb{P}_s[t]^2 = \sum_{t \in RB(s)} \frac{1}{|RB(s)|^2} = \frac{1}{(2s-3)!!} \,.$$

□

The term for the upper bound in Corollary 7 has a simple interpretation in terms of expected values as we show in the following lemma.

Lemma 10 *Let T and T' be randomly generated trees under the uniform model from $RB(n)$. The expected number of pairs of trees from $RB(n)$ with agreement subtrees of size s is equal to*

$$\psi_{n,s} = \binom{n}{s} \frac{1}{(2s-3)!!} .$$

Proof. Treating $X^{(s)}$ in equation (6.2) as a random variable, then taking expected values we obtain

$$\begin{aligned}
\mathbb{E}[X^{(s)}] &= \sum_{S\subseteq[n]:|S|=s} \mathbb{E}[X_S] \\
&= \sum_{S\subseteq[n]:|S|=s} \mathbb{P}[X_S = 1] \\
&= \sum_{S\subseteq[n]:|S|=s} \sum_{t\in RB(s)} \mathbb{P}_s[t]^2 \\
&= \binom{n}{s} \frac{1}{(2s-3)!!} .
\end{aligned}$$

□

As we show in the following proposition, the asymptotic behaviour of $\psi_{n,s}$ as $s \to \infty$ depends on the size of s relative to n.

Proposition 5 *Asymptotically, as $s \to \infty$, with $n \geq s$, we have*

(i) $\psi_{n,s} \leq \sqrt{\frac{s}{\pi}} \left[\frac{ne^2}{2s^2}\right]^s$

(ii) $\psi_{n,s} \geq \sqrt{\frac{s}{\pi}} \left[\frac{(n-s)e^2}{2s^2}\right]^s$

As $s \to \infty$, with $n \leq \frac{2}{e^2}s^2\lambda$ for some constant $\lambda < 1$, then $\psi_{n,s} \to 0$. As $s \to \infty$, with $n \geq \frac{2}{e^2}s^2\lambda$ for some constant $\lambda \geq 1$, then $\psi_{n,s} \to \infty$.

Proof. We have

$$(2s-3)!! = \frac{(2s-3)!}{(s-2)!2^{s-2}} .$$

Therefore

$$\begin{aligned}
\psi_{n,s} &= \frac{n!}{s!(n-s)!}\frac{(s-2)!2^{s-2}}{(2s-3)!} \\
&= \frac{n(n-1)(n-2)\cdots(n-s+1)}{s(s-1)}\frac{2^{s-2}}{(2s-3)!} \\
&\leq \frac{n^s 2^{s-2}}{s(s-1)(2s-3)!} \qquad \text{since } n(n-1)(n-2)\cdots(n-s+1) \leq n^s \\
&= \frac{(2n)^s}{4s(s-1)(2s-3)!} .
\end{aligned} \tag{6.4}$$

Asymptotic approximations and lower and upper bounds for the factorial function are given by (see [24])

$$\sqrt{2\pi}n^{n+1/2}e^{-n}e^{(12n+1)^{-1}} < n! < \sqrt{2\pi}n^{n+1/2}e^{-n}e^{(12n)^{-1}} . \tag{6.5}$$

In particular, since $e^{(12n+1)^{-1}} \geq 1$, then a lower bound is

$$n! > \sqrt{2\pi}n^{n+1/2}e^{-n} .$$

Using this lower bound in (6.4) then

$$\begin{aligned}
\psi_{n,s} &\leq \frac{(2n)^s}{4s(s-1)\sqrt{2\pi}(2s-3)^{2s-3+1/2}e^{-2s+3}} \\
&= \frac{1}{4\sqrt{2\pi}e^3}\frac{(2s-3)^{5/2}}{s(s-1)}\left[\frac{2ne^2}{(2s-3)^2}\right]^s .
\end{aligned}$$

For large s

$$\left[\frac{1}{(2s-3)^2}\right]^s \to e^3\left[\frac{1}{4s^2}\right]^s \qquad \text{and} \qquad \frac{(2s-3)^{5/2}}{s(s-1)} \to 2^{5/2}s^{1/2} .$$

So, asymptotically, an upper bound for $\psi_{n,s}$ is

$$\sqrt{\frac{s}{\pi}}\left[\frac{e^2 n}{2s^2}\right]^s .$$

If the term in square brackets is less than or equal to some constant λ, where $\lambda < 1$, then this upper bound on $\psi_{n,s}$ goes to zero as s goes to infinity. This condition is satisfied provided

$$n \leq \frac{2s^2}{e^2}\lambda .$$

Thus, as $s \to \infty$, with $n \leq \frac{2s^2}{e^2}\lambda$ for some constant $\lambda < 1$, then $\psi_{n,s} \to 0$.

A lower bound may also be found for $\psi_{n,s}$. Starting with, as before, the equation for the mean, we have

$$\begin{aligned}
\psi_{n,s} &= \frac{n(n-1)(n-2)\cdots(n-s+1)}{s(s-1)}\frac{2^{s-2}}{(2s-3)!} \\
&\geq \frac{(n-s+1)^s}{s(s-1)}\frac{2^{s-2}}{(2s-3)!} && \text{since } n(n-1)(n-2)\cdots(n-s+1) \geq (n-s+1)^s \\
&\geq \frac{(n-s)^s}{s(s-1)}\frac{2^{s-2}}{(2s-3)!} && \text{as } (n-s+1)^s \geq (n-s)^s \\
&\geq \frac{(n-s)^s}{s^2}\frac{2^{s-2}}{(2s-3)!} && \text{since } \frac{1}{s(s-1)} \geq \frac{1}{s^2} \\
&= \frac{[2(n-s)]^s}{4s^2(2s-3)!} . && (6.6)
\end{aligned}$$

From the upper bound for the factorial function given in (6.5) it follows that

$$(2s-3)! \leq \sqrt{2\pi}e^3\frac{(2s-3)^{2s}}{(2s-3)^{5/2}}\frac{e^{\frac{1}{24s-36}}}{e^{2s}} ,$$

and subsequently,

$$\frac{1}{(2s-3)!} \geq \frac{1}{\sqrt{2\pi}e^3}\frac{(2s-3)^{5/2}}{(2s-3)^{2s}}\frac{e^{2s}}{e^{\frac{1}{24s-36}}} .$$

Substituting this lower bound into (6.6) then

$$\psi_{n,s} \geq \frac{1}{4\sqrt{2\pi}e^3}\frac{(2s-3)^{5/2}}{s^2}\left[\frac{2(n-s)e^2}{(2s-3)^2}\right]^s \frac{e^{2s}}{e^{\frac{1}{24s-36}}}$$
$$= \frac{1}{4\sqrt{2\pi}e^3}\frac{(2s-3)^{5/2}}{s^2}\frac{1}{e^{\frac{1}{24s-36}}}\left[\frac{(n-s)e^2}{(2s-3)^2}\right]^s .$$

For large s

$$\frac{(2s-3)^{5/2}}{s^2} \to 2^{5/2}s^{1/2} \; ; \quad \left[\frac{1}{(2s-3)^2}\right]^s \to e^3\left[\frac{1}{4s^2}\right]^s \; ; \quad \frac{1}{e^{\frac{1}{24s-36}}} \to 1 .$$

So, asymptotically, a lower bound for $\psi_{n,s}$ is

$$\sqrt{\frac{s}{\pi}}\left[\frac{(n-s)e^2}{2s^2}\right]^s .$$

This lower bound goes to infinity as s becomes large if the term in square brackets is greater than or equal to some constant $\lambda > 1$. This is equivalent to the condition

$$n \geq \frac{2s^2}{e^2}\lambda + s \sim \frac{2s^2}{e^2}\lambda .$$

Thus, as $s \to \infty$, with $n \geq \frac{2s^2}{e^2}\lambda$ for some constant $\lambda > 1$, then $\psi_{n,s} \to \infty$. □

The implications of this proposition are that if the number of leaves on the two trees is large enough ($n \geq 2s^2\lambda/e^2$) then the expected number of MASTs of size s goes to infinity as $s \to \infty$ (see Lemma 10). Tying back in with Corollary 7, then if the number of leaves of the two trees is small enough ($n \leq 2s^2\lambda/e^2$) then the probability of having a MAST of size s or larger goes to zero as $s \to \infty$.

6.2.2 Yule Model

For the Yule model we can derive a recursion for the sum of squared probabilities that occur in Theorem 10.

Proposition 6 *Let $RB(s)$ be the set of rooted labelled trees on s leaves. The summation*

$$S_n = \sum_{t \in RB(s)} \mathbb{P}_s[t]^2$$

satisfies the recursion

$$S_n = \frac{2}{(n-1)^2} \sum_{r=1}^{n-1} \frac{S_r S_{n-r}}{\binom{n}{r}} \qquad \text{where } n \geq 2,\ S_1 = 1\ .$$

If we let $y(x) = \sum_{n=1}^{\infty} a_n x^n$, *where* $a_n = n! S_n$, *then we have*

$$x^2 \frac{d^2y}{dx^2} - x\frac{dy}{dx} + y = 2y^2, \qquad y(0) = 0 \quad y'(0) = 1\ .$$

Proof. We first derive the recursion, then show that $y(x)$ satisfies the stated differential equation. Let t be a labelled tree on n leaves with a left subtree of t_1 of size r, and a right subtree t_2 of size $n - r$:

$$t = t_1 + t_2\ .$$

Let the probability for the tree t on n leaves be $\mathbb{P}_n[t]$. For the Yule model the labelled trees satisfy the recursion (1.15), which upon squaring becomes

$$\mathbb{P}_n[t]^2 = \frac{4}{(n-1)^2} \binom{n}{r}^{-2} \mathbb{P}_r[t_1]^2 \mathbb{P}_{n-r}[t_2]^2\ .$$

So for n odd the sum of squares can be expressed as

$$S_n = \frac{4}{(n-1)^2} \sum_{r<\frac{n}{2}} \sum_{t_1,t_2} \binom{n}{r}^{-2} \mathbb{P}_r[t_1]^2 \mathbb{P}_{n-r}[t_2]^2\ ,$$

where the leaf-labels of the trees t_1, t_2 form a partition of $[n]$.

Separating out the $\binom{n}{r}$ term gives

$$\begin{aligned} S_n &= \frac{4}{(n-1)^2} \sum_{r<\frac{n}{2}} \binom{n}{r}^{-2} \sum_{t_1,t_2} \mathbb{P}_r[t_1]^2 \mathbb{P}_{n-r}[t_2]^2 \\ &= \frac{4}{(n-1)^2} \sum_{r<\frac{n}{2}} \binom{n}{r}^{-1} \sum_{\substack{t_1 \\ \mathcal{L}(t_1)\in[r]}} \mathbb{P}_r[t_1]^2 \sum_{\substack{t_2 \\ \mathcal{L}(t_1)\in[n-r]}} \mathbb{P}_{n-r}[t_2]^2 \\ &= \frac{4}{(n-1)^2} \sum_{r<\frac{n}{2}} \frac{S_r S_{n-r}}{\binom{n}{r}}\ . \end{aligned}$$

So, for n odd, the sum of the squared probabilities satisfies the recursion

$$S_n = \frac{4}{(n-1)^2} \sum_{r<\frac{n}{2}} \frac{S_r S_{n-r}}{\binom{n}{r}} .$$

For n even the additional term for $r = \frac{n}{2}$ is

$$S_e = \frac{1}{2} \frac{4}{(n-1)^2} \sum_{t_1,t_2} \binom{n}{n/2}^{-2} P_{n/2}[t_1]^2 P_{n/2}[t_2]^2 ,$$

where t_1, t_2 each have $\frac{n}{2}$ leaves for which the leaf-labels form a partition of $[1, n]$. Factorising out $\binom{n}{n/2}$ then

$$\begin{aligned} S_e &= \frac{2}{(n-1)^2} \binom{n}{n/2}^{-1} \sum_{\substack{t_1 \\ \mathcal{L}(t_1) \in [n/2]}} \mathbb{P}_{n/2}[t_1]^2 \sum_{\substack{t_2 \\ \mathcal{L}(t_2) \in [n/2]}} \mathbb{P}_{n/2}[t_2]^2 \\ &= \frac{2}{(n-1)^2} \binom{n}{n/2}^{-1} S_{n/2}^2 . \end{aligned}$$

So for n even the sum of the squared probabilities satisfies the recursion

$$S_n = \frac{4}{(n-1)^2} \sum_{r<n/2} \frac{S_r S_{n-r}}{\binom{n}{r}} + \frac{2}{(n-1)^2} \frac{1}{\binom{n}{n/2}} S_{n/2}^2 .$$

The recursions for n odd and n even may be combined to give

$$S_n = \frac{2}{(n-1)^2} \sum_{r=1}^{n-1} \frac{S_r S_{n-r}}{\binom{n}{r}} \qquad \text{where } n \geq 2,\ S_1 = 1.$$

Letting $a_n = n! S_n$ allows the recursion to be rewritten in the simpler form

$$a_n = \frac{2}{(n-1)^2} \sum_{r=1}^{n-1} a_r a_{n-r} . \tag{6.7}$$

We now show that the generating function for the coefficients a_n satisfies a second order ordinary differential equation. Let the generating function for the coefficients a_n be

$$y(x) = \sum_{n=1}^{\infty} a_n x^n .$$

From the generating function we have

$$\sum_{n=1}^{\infty} n a_n x^n = x y'(x) \,, \tag{6.8}$$

$$\sum_{n=1}^{\infty} n^2 a_n x^n = x^2 y''(x) + x y'(x) \,. \tag{6.9}$$

Rewriting (6.7) gives

$$\frac{1}{2}[n^2 - 2n + 1] a_n = \sum_{r=1}^{n-1} a_r a_{n-r} \,.$$

Multiplying both sides by x^n then summing over n gives

$$\frac{1}{2}\sum_{n=1}^{\infty} n^2 a_n x^n - \sum_{n=1}^{\infty} n a_n x^n + \frac{1}{2}\sum_{n=1}^{\infty} a_n x^n = \sum_{n=1}^{\infty}\left[\sum_{r=1}^{n-1} a_r a_{n-r}\right] x^n \,.$$

Recognising the right-hand side as $y(x)^2$ and using the relationships (6.8), (6.9) then

$$\frac{1}{2}x^2 y''(x) - \frac{1}{2}x y'(x) + \frac{1}{2}y(x) = y(x)^2 \,.$$

Multiplying both sides by two, and setting appropriate initial conditions, gives us the second-order, non-linear differential equation satisfied by the generating function

$$x^2 \frac{d^2 y}{dx^2} - x\frac{dy}{dx} + y = 2y^2, \qquad y(0) = 0 \quad y'(0) = 1 \,.$$

□

Chapter 7

The Entropy of Probability Models

7.1 Introduction

Entropy, in the mathematical sense we define in the next section, had its origin in statistical physics. Its uses since then have been expanded by the development of information theory, where entropy has centre place. The concept of entropy now has widespread use in communication theory, computer science, and statistics [19].

Its use in biology, while leading to some interesting developments, has been more limited and seems not to have yielded any deep insights so far. Some early work used entropy as a measure of ecological diversity, where a population has high diversity if it has a large number of species and the number of each species is about the same [55]. Later work has looked at applications to molecular biology (especially the DNA code), and attempts have been made to integrate evolutionary and ecological theory into an information theory framework [12, 77].

In this chapter we calculate exact and asymptotic formulae for the entropy under the comb, uniform and Yule models on labelled rooted trees. Comparing the entropies reveals that the entropy for the Yule model is just under that for the uniform model, while the entropy for the comb model is less than both.

7.2 Entropy

The information content of an event E is

$$I(E) = -\log P(E) ,$$

where the units are *bits* if the logarithm is taken to base 2, and *nats* or *nits* if the base used is e. It follows from this definition that if an event has low probability then it has high information content, the so called 'surprise' interpretation of information [6, pp. 93-94]

Let X be a discrete random variable that can take the values $\{x_1, x_2, \ldots, x_m\}$, with associated probabilities $\{p_1, p_2, \ldots, p_m\}$, and information $\{I(x_1), I(x_2), \ldots, I(x_m)\}$. Then the *entropy* ($\mathbb{J}$) of X is the mean information:

$$\mathbb{J} = \mathbb{E}[I(X)] = -\sum_{j=1}^{m} p_j \log p_j . \tag{7.1}$$

Let T be a random variable whose values are labelled rooted tree on n leaves generated under some random model (e.g. comb, uniform, or Yule). We define $\mathbb{J}_n$ to be the entropy of T.

7.3 The Comb Model

Let T be a labelled rooted tree generated under the comb model.

Theorem 11 *The entropy of T is*

$$\mathbb{J}_n = \log(n!) - \log(2) .$$

Asymptotically we have

$$\mathbb{J}_n - n\log(n) + n \sim -\frac{1}{2}\log(n) ,$$

where the logarithm is to base e.

Proof. Under the comb model the caterpillar shape has probability one, and each associated labelled tree has probability $p_i = 2/n!$. Substituting for p_i in equation (7.1) gives the stated formula for the entropy. The asymptotic result follows from Stirling's asymptotic formula for the factorial [51, p. 1067]. □

7.4 The Uniform Model

Let T be a random variable representing a labelled rooted tree on n leaves generated by the uniform model. Let $RB(n)$ be the set of all possible such trees on n leaves, where the probability of the ith tree in the set is denoted by p_i.

Theorem 12 *The entropy of T is*

$$\mathbb{J}_n = \log |RB(n)| = \log(2n-3)!! \; .$$

Asymptotically we have

$$\mathbb{J}_n - n\log(n) + c_1 n \sim -\log(n) \; ,$$

where $c_1 = 1 - log(2) \approx 0.307$, and the logarithm is to base e throughout.

Proof. The entropy of T is

$$\begin{aligned}
\mathbb{J}_n &= -\sum_{i=1}^{|RB(n)|} p_i \log p_i \\
&= -\sum_{i=1}^{|RB(n)|} \tfrac{1}{|RB(n)|} \log \tfrac{1}{|RB(n)|} \qquad \text{(uniform model)} \\
&= \log |RB(n)| = \log(2n-3)!! \; .
\end{aligned}$$

For the asymptotic approximation first note that

$$|RB(n)| = \frac{(2n-3)!}{(n-2)!2^{n-2}} \; .$$

Thus we have

$$\log(|RB(n)|) = \log(2n-3)! - \log(n-2)! - (n-2)\log(2) \; .$$

Using Stirling's asymptotic approximation [51, p. 1067],

$$\log(x!) - x\log(x) + x - \frac{1}{2}\log(x) \sim \frac{1}{2}\log(2\pi) \; ,$$

and retaining the higher order terms gives the required asymptotic result for the entropy. □

Note that the entropy obtained under the uniform model is the maximum possible entropy for a discrete probability distribution on $|RB(n)|$ states [6, p. 97]. The minimum entropy (of zero) is obtained when one labelled tree has a probability of one, while all the other labelled trees have probability of zero.

7.5 The Yule Model

Before we embark on calculating the entropy for the Yule model we prove the following lemma which will simplify a number of summations we will be required to do.

Lemma 11 *Let t_1, t_2 be leaf-labelled trees on r and $n-r$ leaves respectively, where the leaf sets $\mathcal{L}(t_1)$ and $\mathcal{L}(t_2)$ form a partition of $[n]$. If F and G are arbitrary functions then we have the identity*

$$(i) \sum_{t_1,t_2} F(p_{t_1}^{(r)}) G(p_{t_2}^{(n-r)}) = \binom{n}{r} \sum_{\mathcal{L}(t_1)\in[1,r]} F(p_{t_1}^{(r)}) \sum_{\mathcal{L}(t_2)\in[r+1,n]} G(p_{t_2}^{(n-r)}) ,$$

and when F and G are the identity functions we have

$$(ii) \sum_{t_1,t_2} p_{t_1}^{(r)} p_{t_2}^{(n-r)} = \binom{n}{r} .$$

Proof. Part (i) follows from invariance of tree probabilities under relabelling. Since tree probabilities sum to one, then when F and G are the identity functions we get part (ii). □

Theorem 13 *Let T be a random variable representing a labelled rooted tree on n leaves generated by the Yule model. Let the entropy of T be $\mathbb{J}_n$. We have the recursion for $n \geq 3$:*

$$\mathbb{J}_n = \log \frac{n-1}{2} + \frac{2}{n-1} \sum_{k=1}^{n-1} J_k + \frac{1}{n-1} \sum_{k=1}^{n-1} \log \binom{n}{k}, \qquad J_2 = 0 .$$

This recursion has the explicit solution for $n \geq 3$:

$$\mathbb{J}_n = n \sum_{k=2}^{n-1} \frac{g(k)}{k+1} ,$$

where $g(k) = \frac{1-k}{k}\log\frac{k-1}{2} + \log\frac{k}{2} + \log(k+1) - \frac{1}{k}\log k!$. *Asymptotically we have*

$$J_n - n\log(n) + c_1 n \sim -\frac{1}{2}\log(n)\,,$$

where $c_1 = \log(2)\log(\frac{200}{49e}) + \log(9)\log(7/10) + 2\,\mathrm{Li}_2(7/4) - 2\,\mathrm{Li}_2(5/2) - 1 \approx 0.493$, *and* $\mathrm{Li}_2(x) = \int_1^x \frac{\log t}{1-t}dt$, *with the logarithm taken to base e throughout.*

Proof. We first derive separate recursions for $\mathbb{J}_n$ for n odd and n even. We then combine these separate recursions into a single recursion, which we solve to get the explicit solution.

Starting from the basic definition of entropy, then as before we have

$$\mathbb{J}_n = -\sum_{i=1}^{|RB(n)|} p_{t_i}^{(n)} \log p_{t_i}^{(n)}\,. \tag{7.2}$$

A useful shorthand notation, valid for n odd, is

$$\sum_{\text{all trees}} \equiv \sum_{r<\frac{n}{2}} \sum_{t_1,t_2}$$

where r is a positive integer, t_1, t_2 are labelled trees on r and $n-r$ leaves respectively, and the leaf-labels of t_1 and t_2 form a partition of $[n]$.

Substituting in (7.2) using the recursive formula (1.15) for probabilities under the Yule model gives

$$\mathbb{J}_n = -\sum_{\text{all trees}} \frac{2}{n-1}\binom{n}{r}^{-1} p_{t_1}^{(r)} p_{t_2}^{(n-r)} \log\left\{\frac{2}{n-1}\binom{n}{r}^{-1} p_{t_1}^{(r)} p_{t_2}^{(n-r)}\right\}\,. \tag{7.3}$$

Expanding out the log term and factorising the $n-1$ term gives

$$\begin{aligned}\mathbb{J}_n = \frac{-2}{n-1}\Big\{ &\sum_{\text{all trees}} \tbinom{n}{r}^{-1} p_{t_1}^{(r)} p_{t_2}^{(n-r)} \log\tfrac{2}{n-1} + \sum_{\text{all trees}} \tbinom{n}{r}^{-1} p_{t_1}^{(r)} p_{t_2}^{(n-r)} \log \tbinom{n}{r}^{-1} \\ &+ \sum_{\text{all trees}} \tbinom{n}{r}^{-1} p_{t_1}^{(r)} p_{t_2}^{(n-r)} \log p_{t_1}^{(r)} + \sum_{\text{all trees}} \tbinom{n}{r}^{-1} p_{t_1}^{(r)} p_{t_2}^{(n-r)} \log p_{t_2}^{(n-r)}\Big\}\,.\end{aligned}$$

Or in a more compact form

$$\mathbb{J}_n = \frac{-2}{n-1}\Big\{A + B + C + D\Big\}\,. \tag{7.4}$$

For the separate terms A, B, C, D we have

$$\begin{aligned} A &= \log \frac{2}{n-1} \sum_{\text{all trees}} \binom{n}{r}^{-1} p_{t_1}^{(r)} p_{t_2}^{(n-r)} \\ &= \log \frac{2}{n-1} \sum_{r<\frac{n}{2}} \binom{n}{r}^{-1} \sum_{t_1,t_2} p_{t_1}^{(r)} p_{t_2}^{(n-r)} \\ &= \frac{1}{2}(n-1) \log \tfrac{2}{n-1} \qquad \text{(Lemma 11)} , \end{aligned}$$

$$\begin{aligned} B &= \sum_{\text{all trees}} \tbinom{n}{r}^{-1} p_{t_1}^{(r)} p_{t_2}^{(n-r)} \log \tbinom{n}{r}^{-1} \\ &= \sum_{r<\frac{n}{2}} \tbinom{n}{r}^{-1} \log \tbinom{n}{r}^{-1} \sum_{t_1,t_2} p_{t_1}^{(r)} p_{t_2}^{(n-r)} \\ &= \sum_{r<\frac{n}{2}} \log \tbinom{n}{r}^{-1} \qquad \text{(Lemma 11)} , \end{aligned}$$

$$\begin{aligned} C &= \sum_{\text{all trees}} \tbinom{n}{r}^{-1} p_{t_1}^{(r)} p_{t_2}^{(n-r)} \log p_{t_1}^{(r)} \\ &= \sum_{r<\frac{n}{2}} \tbinom{n}{r}^{-1} \sum_{t_1,t_2} p_{t_2}^{(n-r)} p_{t_1}^{(r)} \log p_{t_1}^{(r)} \\ &= \sum_{r<\frac{n}{2}} \sum_{\mathcal{L}(t_1)\in[1,r]} p_{t_1}^{(r)} \log p_{t_1}^{(r)} \qquad \text{(Lemma 11)} \\ &= -\sum_{r<\frac{n}{2}} \mathbb{J}_r(T) . \end{aligned}$$

Likewise, using symmetry,

$$D = -\sum_{r<\frac{n}{2}} \mathbb{J}_{n-r}(T) .$$

Combining these terms back together in (7.4) gives

$$\mathbb{J}_n = \log \tfrac{n-1}{2} + \tfrac{2}{n-1} \sum_{r<\frac{n}{2}} \left[\mathbb{J}_r + \mathbb{J}_{n-r} + \log \tbinom{n}{r} \right] .$$

In fact

$$\sum_{r<\frac{n}{2}} [\mathbb{J}_r + \mathbb{J}_{n-r}] = \sum_{r<\frac{n}{2}} \mathbb{J}_r + \sum_{r>\frac{n}{2}} \mathbb{J}_r = \sum_{k=1}^{n-1} \mathbb{J}_r \ ,$$

and similarly

$$\sum_{r<\frac{n}{2}} \log \binom{n}{r} = \frac{1}{2} \sum_{k=1}^{n-1} \log \binom{n}{k} \ .$$

So for n odd the entropy is given by the following recursion

$$\mathbb{J}_n = \log \frac{n-1}{2} + \frac{2}{n-1} \sum_{k=1}^{n-1} \mathbb{J}_k + \frac{1}{n-1} \sum_{k=1}^{n-1} \log \binom{n}{k} \ . \tag{7.5}$$

For n even an additional term $\mathbb{J}_e$ is added to the right-hand side of (7.3):

$$\mathbb{J}_e = -\frac{1}{2} \sum_{r=\frac{n}{2}} \sum_{t_1,t_2} \frac{2}{n-1} \binom{n}{r}^{-1} p_{t_1}^{(r)} p_{t_2}^{(n-r)} \log \left\{ \frac{2}{n-1} \binom{n}{r}^{-1} p_{t_1}^{(r)} p_{t_2}^{(n-r)} \right\} \ ,$$

where the factor of $1/2$ takes into account symmetry. Expanding out the log term and factorising the parts that depend only on n gives

$$\begin{aligned} \mathbb{J}_e = & -\frac{1}{n-1} \binom{n}{n/2}^{-1} \log \left[\frac{2}{n-1} \binom{n}{n/2}^{-1} \right] \sum_{t_1,t_2} p_{t_1}^{(\frac{n}{2})} p_{t_2}^{(\frac{n}{2})} \\ & - \frac{1}{n-1} \binom{n}{n/2}^{-1} \sum_{t_1,t_2} p_{t_1}^{(\frac{n}{2})} p_{t_2}^{(\frac{n}{2})} \log p_{t_1}^{(\frac{n}{2})} \\ & - \frac{1}{n-1} \binom{n}{n/2}^{-1} \sum_{t_1,t_2} p_{t_1}^{(\frac{n}{2})} p_{t_2}^{(\frac{n}{2})} \log p_{t_2}^{(\frac{n}{2})} \ . \end{aligned}$$

Using Lemma 11 repeatedly gives

$$\begin{aligned} \mathbb{J}_e = & -\frac{1}{n-1} \binom{n}{n/2}^{-1} \binom{n}{n/2} \log \left[\frac{2}{n-1} \binom{n}{n/2}^{-1} \right] \\ & - \frac{1}{n-1} \sum_{\mathcal{L}(t_1) \in [1,r]} p_{t_1}^{(\frac{n}{2})} \log p_{t_1}^{(\frac{n}{2})} \\ & - \frac{1}{n-1} \sum_{\mathcal{L}(t_2) \in [r+1,n]} p_{t_2}^{(\frac{n}{2})} \log p_{t_2}^{(\frac{n}{2})} \ . \end{aligned}$$

Identifying the summation terms as entropy terms gives as the additional part for n even

$$\mathbb{J}_e = \frac{1}{n-1}\log\frac{n-1}{2} + \frac{2}{n-1}\mathbb{J}_{n/2} + \frac{1}{n-1}\log\binom{n}{n/2} .$$

So for n even we have

$$\mathbb{J}_n = \log\tfrac{n-1}{2} + \tfrac{2}{n-1}\sum_{r<\frac{n}{2}}\left[\mathbb{J}_r + \mathbb{J}_{n-r} + \log\tbinom{n}{r}\right] + \mathbb{J}_e .$$

Since this equation is the same as (7.5) when n is even then we have

$$\mathbb{J}_n = \log\frac{n-1}{2} + \frac{2}{n-1}\sum_{k=1}^{n-1}\mathbb{J}_k + \frac{1}{n-1}\sum_{k=1}^{n-1}\log\binom{n}{k} \qquad n \text{ odd or even.}$$

We now solve this recursion. Let $f(n) = \log\frac{n-1}{2} + \frac{1}{n-1}\sum_{k=1}^{n-1}\log\binom{n}{k}$ then the recursion can be written as

$$\mathbb{J}_n = \frac{2}{n-1}\sum_{k=1}^{n-1}\mathbb{J}_k + f(n) .$$

Substituting for the term $\mathbb{J}_n$, in the expression for $\mathbb{J}_{n+1}$, leads to the form

$$\mathbb{J}_{n+1} = \frac{n+1}{n}\mathbb{J}_n + g(n) ,$$

where

$$\begin{aligned} g(n) &= f(n+1) - \frac{n-1}{n}f(n) \\ &= \frac{1-n}{n}\log\frac{n-1}{2} + \log\frac{n}{2} + \log(n+1) - \frac{1}{n}\log n! . \end{aligned}$$

The explicit solution to this is [74, p. 233]

$$\begin{aligned} \mathbb{J}_n &= \left(\prod_{k=2}^{n-1}\frac{k+1}{k}\right)C + \sum_{m=2}^{n-2}\left(\prod_{k=m+1}^{n-1}\frac{k+1}{k}\right)g(m) + g(n-1) \\ &= n\sum_{m=2}^{n-1}\frac{g(m)}{m+1} \qquad \text{where } C = 0 \text{ since } \mathbb{J}_3 = \log 3 . \end{aligned} \tag{7.6}$$

An asymptotic approximation to the explicit formula is obtained by deriving an exact formula for one of the terms in the summation, and using integral approximations for the rest. The log function is taken to base e for all the integral approximations. Firstly, the summation in the explicit formula can be expanded as

$$\sum_{k=2}^{n-1} \frac{g(k)}{k+1} = -\log(2) \sum_{k=2}^{n-1} \frac{1}{k(k+1)} + \sum_{k=2}^{n-1} \frac{\log(k+1)}{k+1} + \sum_{k=2}^{n-1} \frac{\log(k)}{k+1} + \sum_{k=2}^{n-1} \left[\frac{\log(k-1)}{k(k+1)} - \frac{\log(k-1)}{k+1} \right] + \sum_{k=2}^{n-1} \frac{\log k!}{k(k+1)} . \tag{7.7}$$

For the first term in (7.7) we have the exact result

$$-\log(2) \sum_{k=2}^{n-1} \frac{1}{k(k+1)} = \left(\frac{1}{n} - \frac{1}{2} \right) \log(2) .$$

For the second term we use the integral approximation

$$\sum_{k=2}^{n-1} \frac{\log(k+1)}{k+1} \approx \int_{5/2}^{n+1/2} \frac{\log(x)dx}{x} .$$

Likewise for the third term

$$\sum_{k=2}^{n-1} \frac{\log(k)}{k+1} \approx \int_{3/2}^{n-1/2} \frac{\log(x)dx}{x+1} .$$

Rewriting the fourth term and using an integral approximation gives

$$\sum_{k=2}^{n-1} \left[\frac{\log(k-1)}{k(k+1)} - \frac{\log(k-1)}{k+1} \right] \approx \int_{3/2}^{n-1/2} \frac{\log(x)dx}{x+1} - 2 \int_{3/2}^{n-3/2} \frac{\log(x)}{x+2} - \frac{1}{n} \log(n-1) .$$

Expanding out the $\log k!$ term in the fifth term, and changing the order of summation gives

$$\sum_{k=2}^{n-1} \frac{\log k!}{k(k+1)} = \sum_{i=2}^{n-1} \frac{\log(i)}{i} - \frac{1}{n} \log(n-1)! .$$

Using an integral approximation for the summation term on the right-hand side, and an asymptotic approximation for the factorial, then we have for the fifth term

$$-\sum_{k=2}^{n-1} \frac{\log k!}{k(k+1)} \approx \log(n-1) - \frac{1}{2n}\log(n-1) - 1 + \frac{1}{2n}[2+\log(2\pi)] - \int_{3/2}^{n-1/2} \frac{\log(x)dx}{x} .$$

Substituting the five approximation terms into (7.6), then retaining only the higher order terms, gives the required asymptotic approximation. □

7.6 Discussion

It is interesting to compare the entropy of the comb and Yule models to that of the uniform, remembering that the entropy for the uniform model is the maximum possible for a discrete probability distribution. The entropy of the Yule model is in fact just under that for the uniform model, while the entropy of the comb model is distinctly less than the Yule model (Figure 7.1). This uniform-like behaviour of the Yule model is not entirely unexpected. In the Yule model labelled trees are given probability uniformly over a given tree shape, so a large element of uniformity is still present. In addition, this uniform element increases as n increases. For example, the fully symmetric tree shape on $n = 2^k$ leaves has the least number of labelled trees associated with it of all tree shapes ($n!/2^{n-1}$), but this number still grows rapidly with n. Asymptotically the comb, Yule and uniform models all have an entropy of $n\log_e(n)$, indicating that the aspect of uniformity present in the labelling of trees dominates for large n. Interpreting the entropy as the mean information, then specifying that the tree probabilities follow a Yule distribution leads to little decrease in the mean information compared to the uniform model.

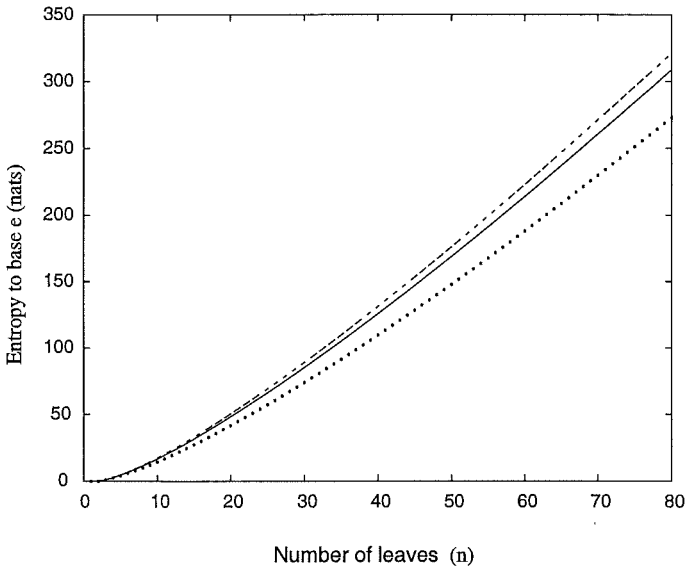


Figure 7.1: Entropy for the uniform model (dashed line) , Yule model (solid line), and comb model (dotted line).

Chapter 8

Group Elimination

8.1 Introduction

In this chapter we introduce and investigate a property of probability distributions on rooted trees called *group elimination*, of which a special case is the *sampling consistency* property. First we introduce some motivation and terminology for this property (Section 8.2). Next we determine the type and structure of the equations implied by group elimination (Section 8.3). We then show that the comb, uniform, and Yule models satisfy the group elimination property (Section 8.4). An unresolved conjecture is that these are the only three models that satisfy group elimination. Lastly, we show that if a probability distribution on trees satisfies sampling consistency then there is an upper bound for the probability of the fully symmetric tree shapes (Section 8.5).

8.2 Motivation and Terminology

The property of group elimination for probability distributions on rooted trees was introduced by Aldous by analogy with the mathematical theory for neutral population genetics [4]. Neutral population genetics is concerned with changes in allele (different versions of a gene) frequency that are due to random effects, as opposed to changes in allele frequency due to selection effects. A fundamental formula in neutral population genetics, with mutation included, is the Ewans sampling formula for the probability distribution of the number and types of alleles in a sample. This formula satisfies the familiar property of *exchangeability* and the less familiar property of *group elimination.* Transporting the

scene to that of macroevolution, Aldous suggested that by analogy a neutral stochastic model for evolutionary trees should also satisfy these properties. We now give a formal definition of the group elimination property.

Recall that if a set of leaves are the only descendants of some internal node then we say that they form a *group* (also called a *clade* or *cluster*). Let T be a random variable that can take as its value any labelled rooted tree on n leaves, and suppose t is a particular labelled rooted tree on k leaves. A probability distribution on labelled rooted trees has the *group elimination* property if

$$\mathbb{P}_n[T_{|\{1,2,...k\}} = t \mid \{k+1, k+2, ..., n\} \text{ is a group on T }] = \mathbb{P}_k[t] \quad \forall\, n, k, t:\ k < n\,. \tag{8.1}$$

A related property, which we show later to be a special case of group elimination, is *sampling consistency*. A probability distribution on labelled rooted trees has the *sampling consistency* property if

$$\mathbb{P}_n[T_{|\{1,2,...,n-1\}} = t\,] = \mathbb{P}_{n-1}[t]\,, \tag{8.2}$$

where t is a labelled tree on $n-1$ leaves.

8.3 Some Elaboration

In this section we expand on the basic definition of group elimination given in (8.1). More specifically, if tree probabilities satisfy group elimination and exchangeability we show that this implies that they satisfy a set of linear and quadratic equations. First some further definitions. Let P_n^j be the probability of the jth tree shape on n leaves, where the tree shape has symmetry index σ_j (refer back to Figure 1.2). Suppose t_k^q is a particular labelled tree on k leaves with shape q, then let $N_n^j(k, t_k^q)$ be the number of labelled trees on n leaves with the jth shape for which $\{k+1, k+2, \ldots, n\}$ form a group and $T_{|\{1,2,\ldots,k\}} = t_k^q$. Also, let $N_n^j(k)$ be the number of labelled trees on n leaves with the jth tree shape for which $\{k+1, k+2, \ldots, n\}$ form a group.

Firstly, using basic conditional probability, we can rewrite the group elimination property as

$$\mathbb{P}_n[\, T_{|\{1,2,\ldots,k\}} = t \text{ and } \{k+1, k+2, \ldots, n\} \text{ is a group }] = \mathbb{P}_k[t]\, \mathbb{P}_n[\, \{k+1, k+2, \ldots, n\} \text{ is a group }] \,. \quad (8.3)$$

We can rewrite the left-hand side of (8.3) as

$$\sum_{j=1}^{\mathcal{S}(n)} N_n^j(k, t_k^q) \frac{2^{\sigma_j}}{n!} P_n^j \,, \quad (8.4)$$

where $\mathcal{S}(n)$ is the number of tree shapes on n leaves. The right-hand side is

$$\frac{2^{\sigma_q} P_k^q}{k!} \sum_{j=1}^{\mathcal{S}(n)} N_n^j(k) \frac{2^{\sigma_j}}{n!} P_n^j \,. \quad (8.5)$$

When we use (8.4) and (8.5) in (8.3) we get the following set of equations for the tree shape probabilities P_n^j.

$$\sum_{j=1}^{\mathcal{S}(n)} N_n^j(k, t_k^q) \frac{2^{\sigma_j}}{n!} P_n^j = \frac{2^{\sigma_q} P_k^q}{k!} \sum_{j=1}^{\mathcal{S}(n)} N_n^j(k) \frac{2^{\sigma_j}}{n!} P_n^j \quad \forall n, k, q\colon\ k < n,\ q = 1, 2, \ldots, \mathcal{S}(k) \,.$$

To further emphasise the structure of the equations we substitute $L_n^j = \frac{2^{\sigma_j}}{n!} P_n^j$, the probability of a labelled tree on n leaves with shape j, to obtain

$$\sum_{j=1}^{\mathcal{S}(n)} N_n^j(k, t_k^q) L_n^j = L_k^q \sum_{j=1}^{\mathcal{S}(n)} N_n^j(k) L_n^j \quad \forall n, k, q\colon\ k < n,\ q = 1, 2, \ldots, \mathcal{S}(k) \,. \quad (8.6)$$

We do not need to add separately the condition that the terms L_k^q must sum to one. This is because this condition is implicit in the group elimination equations, as we show in the following lemma.

Lemma 12 *The group elimination equations* (8.6) *are a sufficient condition for the terms* L_k^q *to sum to one.*

Proof. From the group elimination equations we have

$$L_k^q = \frac{\sum_{j=1}^{\mathcal{S}(n)} N_n^j(k, t_k^q) L_n^j}{\sum_{j=1}^{\mathcal{S}(n)} N_n^j(k) L_n^j} \,.$$

Summing both sides over all labelled trees $t_k \in RB(k)$ then we have

$$\sum_{t_k \in RB(k)} L_k^q = \frac{\sum_{j=1}^{\mathcal{S}(n)} L_n^j \sum_{t_k \in RB(k)} N_n^j(k, t_k^q)}{\sum_{j=1}^{\mathcal{S}(n)} N_n^j(k) L_n^j} = \frac{\sum_{j=1}^{\mathcal{S}(n)} N_n^j(k) L_n^j}{\sum_{j=1}^{\mathcal{S}(n)} N_n^j(k) L_n^j} = 1 \, .$$

□

We now mention some generalities regarding the number of equations at the nth level. Firstly, the equations make no constraints on the tree shape probabilities for $n = 2, 3$. Since there is only one tree shape for either $n = 2$ or $n = 3$ we assign these tree shapes probability one, then use exchangeability to assign the probability for the corresponding labelled trees. Secondly, for $n \geq 5$ and $k \leq 3$, there are no constraints on tree probabilities, as is shown in the following lemma.

Lemma 13 *For $n \geq 5$ and $k \leq 3$ the equations (8.6) yield the trivial equality $0 = 0$.*

Proof. We prove this for $k = 3$, the argument for $k = 1, 2$ is very similar. Let $k = 3$ in (8.6) then we have

$$\sum_{j=1}^{\mathcal{S}(n)} N_n^j(3, t_3^1) L_n^j = \frac{1}{3} \sum_{j=1}^{\mathcal{S}(n)} N_n^j(k) L_n^j \, . \tag{8.7}$$

There are three ways in which we can label the tree shape on three leaves, then attach the group $\{4, 5, \ldots, n\}$, yielding the set of trees on n leaves for which $\{4, 5, \ldots, n\}$ is a group. Of the labelled trees shapes on three leaves only one of them is the tree t_3^1 therefore we have $N_n^j(k) = 3 \times N_n^j(3, t_3^1)$. Substituting for $N_n^j(k)$ in (8.7) yields the trivial equality. □

As we have a non-trivial equation for each labelled tree shape t_k^q ($k \geq 4$) then, at the nth level, we potentially have $\mathcal{F}(n) = \mathcal{S}(4) + \ldots + \mathcal{S}(n-1)$ equations in the $\mathcal{S}(n)$ probabilities L_n^j. In addition, there is from Lemma 12 an implicit condition that will be satisfied automatically, that the probabilities sum to one. As some of the equations may be dependent (as indeed turns out to be the case), then in fact $\mathcal{F}(n)$ is an upper bound on the number of equations. Also, since we are dealing with probabilities we have the further constraint that $L_n^j \geq 0$.

The equations from (8.6) are mostly quadratic equations, but under certain special conditions they are linear. If $k = n - 1$ (sampling consistency) then the right-hand side of

(8.6) becomes L_k^q and the equations are linear. Similarly, if the probabilities L_k^q are known up to $q = n-1$ then (8.6) becomes a linear set of equations in the L_n^j.

It is instructive to explicitly show the equations implied by group elimination for some small values of n. Here we look at the cases for $n = 4, 5, 6$ and the probabilities of the relevant tree shapes for these values of n. We represent the rooted tree shapes by the 'dictionary' notation (see p. 5). For the probabilities of the shapes we let $x = \mathbb{P}[4_1]$, $y = \mathbb{P}[4_2]$, $p_i = \mathbb{P}[5_i]$ (where $i = 1, 2, 3$) and $k_i = \mathbb{P}[6_i]$ (where $i = 1, \ldots, 6$) .

At the $n = 4$ level we have just one equation from $k = 3$:

$$x + y = 1 .$$

Other values of k only give trivial equalities, so for $n = 4$ we only have one equation for the probabilities.

For $n = 5$ we could have up to $\mathcal{F}(5) = 2$ equations. Letting $k = 4$ we get two linear equations:

$$\begin{aligned} 5p_1 + 4p_2 + 2p_3 &= 5x \\ p_2 + 3p_3 &= 5y . \end{aligned}$$

For $n = 6$ we could have up to $\mathcal{F}(6) = 5$ equations. For $n = 6$ and $k = 5$ we get the three linear equations:

$$\begin{aligned} 3k_1 + 2k_2 + k_3 + k_4 &= 3p_1 \\ 2k_2 + 3k_3 + 2k_5 &= 6p_2 \\ k_3 + 4k_4 + 4k_5 + 6k_6 &= 6p_3 . \end{aligned}$$

For $n = 6$ and $k = 4$ we get just one quadratic equation (since $x + y = 1$ gives the other)

$$(1-x)k_1 + (2-2x)k_2 + (2-2x)k_3 + (1-2x)k_4 - 3xk_5 + (2-2x)k_6 = 0 . \qquad (8.8)$$

8.4 Distributions Satisfying Group Elimination

We now show that the comb model, uniform model, and Yule model all satisfy group elimination.

Proposition 7 *The comb distribution on rooted trees satisfies the group elimination property.*

Proof. Suppose T is a labelled tree on n leaves, and let t be a labelled tree on k leaves. If the tree t is not a fully unbalanced tree then the probability of obtaining t under the comb model is zero, so henceforth only the case of the fully unbalanced tree is considered. The expression

$$\mathbb{P}_n[T_{|\{1,2,\ldots,k\}} = t \mid \{k+1, k+2, \ldots, n\} \text{ form a group}]$$

may be rewritten as

$$\frac{\mathbb{P}_n[T_{|\{1,2,\ldots,k\}} = t \text{ and } \{k+1, k+2, \ldots, n\} \text{ form a group}]}{\mathbb{P}_n[\{k+1, k+2, \ldots, n\} \text{ form a group}]} = \frac{N}{D} . \tag{8.9}$$

The numerator is

$$N = \sum_{T:T\in\text{comb}} \mathbb{P}_n[T] \qquad \text{where } T : T_{|\{1,2,\ldots,k\}} = t \text{ and } \{k+1, k+2, \ldots, n\} \text{ form a group.}$$

The probability of a labelled comb tree is $2/n!$, and this probability is the same for all the labelled combs on n leaves, so we have

$$N = \frac{2}{n!} \left[|T_{\text{comb}}| : T_{\text{comb}\,|\{1,2,3,\ldots,k\}} = t \text{ and } \{k+1, k+2, \ldots, n\} \text{ form a group} \right] .$$

The number of comb trees for which the leaves $\{k+1, k+2, \ldots, n\}$ form a group and $T_{\text{comb}|\{1,2,3,\ldots,k\}} = t$ is equal to $2 \times (n-k)!/2$, where the initial factor of two allows for the swapping of the two edges which are adjacent in the tree t, but are not when t is embedded in T. We therefore have

$$N = \frac{2(n-k)!}{n!} .$$

The denominator is

$$\begin{aligned} D &= \sum_{T_{comb}} \mathbb{P}_n[T_{comb}] \quad \text{where } T_{comb} : \{k+1, k+2, ..., n\} \text{ form a group} \\ &= \frac{2}{n!} \left[|T_{comb}| : \{k+1, k+2, ..., n\} \text{ form a group}\right] . \end{aligned}$$

The number of ways of labelling a comb with $(n-k)$ leaves is $(n-k)!/2$, and the remaining k leaves may be labelled in $k!$ ways, so we have

$$\begin{aligned} D &= \frac{2}{n!} \cdot k! \cdot \frac{(n-k)!}{2} \\ &= \frac{k!(n-k)!}{n!} . \end{aligned}$$

Upon substituting back into (8.9) we obtain

$$\begin{aligned} \mathbb{P}_n[T_{|\{1,2,...,k\}} = t \mid \{k+1, k+2, ..., n\} \text{ form a group}] &= \frac{2(n-k)!}{k!(n-k)!} \\ &= \frac{2}{k!} \\ &= \mathbb{P}_k[t] . \end{aligned}$$

□

Proposition 8 *The uniform distribution on rooted trees satisfies the group elimination property.*

Proof. The expression

$$\mathbb{P}_n[T_{|\{1,2,...,k\}} = t \mid \{k+1, k+2, ..., n\} \text{ form a group}]$$

may be rewritten as

$$\frac{\mathbb{P}_n[T_{|\{1,2,...,k\}} = t \text{ and } \{k+1, k+2, ..., n\} \text{ form a group}]}{\mathbb{P}_n[\{k+1, k+2, ..., n\} \text{ form a group}]} . \tag{8.10}$$

Consider the numerator of this expression. The numerator probability may be found by finding the number of trees on n leaves for which $T_{|\{1,2,...,k\}} = t$ and $\{k+1, k+2, ..., n\}$ form a group, then dividing this by the number of labelled trees on n leaves. A subtree with the labels $\{k+1, k+2, ..., n\}$ may be added to the tree t on k leaves in $2k-1$ places, giving a tree on n leaves. Furthermore, the number of subtrees labelled $\{k+1, k+2, ..., n\}$ is $(2n-2k-3)!!$. So the number of trees on n leaves for which $T_{|\{1,2,...k\}} = t$ and $\{k+1, k+2, ...n\}$ form a group is $(2k-1)(2n-2k-3)!!$. This gives

$$\mathbb{P}_n[T_{|\{1,2,...,k\}} = t \text{ and } \{k+1, k+2, ..., n\} \text{ form a group}] = \frac{(2k-1)(2n-2k-3)!!}{(2n-3)!!} .$$

Now consider the denominator of (8.10). There are $(2k-3)!!$ labelled trees on k leaves. A subtree with the labels $\{k+1, k+2, ..., n\}$ may be added to any such tree, giving a tree on n leaves, in $2k-1$ places. Furthermore, the number of subtrees labelled $\{k+1, k+2, ..., n\}$ is $(2n-2k-3)!!$. So the number of trees on n leaves for which $\{k+1, k+2, ..., n\}$ form a group is $(2k-3)!!(2k-1)(2n-2k-3)!! = (2k-1)!!(2n-2k-3)!!$. This gives

$$\mathbb{P}_n[\{k+1, k+2, ..., n\} \text{ form a group}] = \frac{(2k-1)!!(2n-2k-3)!!}{(2n-3)!!} .$$

Therefore, upon substituting back into (8.10) we obtain

$$\begin{aligned} \mathbb{P}_n[T_{|\{1,2,...,k\}} = t \mid \{k+1, k+2, ..., n\} \text{ form a group}] &= \frac{(2k-1)(2n-2k-3)!!}{(2k-1)!!(2n-2k-3)!!} \\ &= \frac{1}{(2k-3)!!} \\ &= \mathbb{P}_k[t] . \end{aligned}$$

Proposition 9 *The Yule distribution on rooted trees satisfies the group elimination property.*

Proof. The expression

$$\mathbb{P}_n[T_{|\{1,2,...,k\}} = t \mid \{k+1, k+2, ..., n\} \text{ form a group}]$$

may be rewritten as

$$\frac{\mathbb{P}_n[T|_{\{1,2,\ldots,k\}} = t \text{ and } \{k+1, k+2, \ldots n\} \text{ form a group}]}{\mathbb{P}_n[\{k+1, k+2, \ldots, n\} \text{ form a group}]}. \tag{8.11}$$

First consider the denominator of this expression. This probability may be found by finding the number of histories for which $\{k+1, k+2, \ldots, n\}$ form a group, then dividing this by the number of possible histories on n leaves. A labelled tree T on n leaves containing $\{k+1, k+2, \ldots, n\}$ as a group can be split into two rooted trees G and G', where G' is the tree for which the leaves $\{k+1, k+2, \ldots, n\}$ form a group and G is the rest of the tree. G' contains $n-k$ leaves so has H_{n-k} histories, while G contains k leaves so has H_k histories.

Let G' be attached in the horizon (see Figure 1.5) between the internodes $i-1$ and i of G. There are i segments in G in the horizon between the internodes $i-1$ and i $(2 \leq i \leq k-1)$. Below the point at which G' attaches to G there are $k-i$ internodes for G. The $k-i$ internodes on G below the point of attachment can be rearranged relative to the $n-k-1$ internodes of G' in $w_{n-i+1,n-k}$ ways, where w_{a_i,b_i} is the number of distinct ways of placing $a_i - b_i - 1$ objects into b_i partitions [13].

So the number of histories in which G' can be attached to G between the internodes $i-1$ and i is $i \cdot H_{n-k} H_k w_{n-i+1,n-k}$. Summing over all horizons gives the number of histories for which the subtree G' is attached to the subtree G as

$$H_{n-k} H_k \sum_{i=1}^{k} i \cdot w_{n-i+1,n-k} \cdot$$

Therefore we have

$$\mathbb{P}_n[\{k+1, k+2, \ldots n\} \text{ form a group}] = \frac{H_{n-k} H_k \sum_{i=1}^{k} i \cdot w_{n-i+1,n-k}}{\frac{n!(n-1)!}{2^{n-1}}} .$$

Now consider the numerator of expression (8.11). Let T be split into two rooted trees t and G', where t is a particular labelled tree on the leaves $\{1, 2, \ldots, k\}$ and G' is the tree for which the leaves $\{k+1, k+2, \ldots, n\}$ form a group. The number of histories in which G' is attached to t between the nodes $i-1$ and i is $i \cdot H_{n-k} H_k(t) w_{n-i+1,n-k}$, where $H_k(t)$ is the number of histories for the tree t. Summing over all the horizons gives the number of histories for which group G' is attached to the subtree t as

$$H_{n-k}H_k(t)\sum_{i=1}^{k} i \cdot w_{n-i+1,n-k} \cdot$$

Therefore

$$\mathbb{P}_n[T_{|\{1,2,...k\}} = t \text{ and } \{k+1, k+2, ...n\} \text{ form a group}] = \frac{H_{n-k}H_k(t)\sum_{i=1}^{k} i \cdot w_{n-i+1,n-k}}{\frac{n!(n-1)!}{2^{n-1}}} .$$

Upon taking the ratio of the numerator and denominator terms we get,

$$\begin{aligned} \mathbb{P}_n[T_{|\{1,2,...k\}} = t \mid \{k+1, k+2, ...n\} \text{ form a group}] &= \frac{H_k(t)}{H_k} \\ &= \mathbb{P}_k(t) . \end{aligned}$$

□

It has been conjectured by Aldous [4] that the Yule, uniform, and comb distributions are the only three distributions that satisfy the group elimination property. We have been unable to prove or disprove this conjecture. An obvious approach to try to find another distribution that satisfies group elimination is take convex combinations of the three that do (i.e. $\alpha_1 P_{Yule} + \alpha_2 P_{uni} + \alpha_3 P_{comb} : \alpha_1 + \alpha_2 + \alpha_3 = 1$). A convex combination automatically satisfies the linear equations for group elimination, but it only seems to satisfy the quadratic equations if all of α_i are zero except for one, a suggestive result but certainly not conclusive.

Another approach is to define a probabilistic edge-adding model that interpolates between the Yule, uniform, and comb models. In the Yule model the next edge is always added to a pendant edge, while in the uniform model the next edge can be added to both internal and pendant edges. The comb model can be defined as that model in which the next edge is always added to an internal edge. If we define parameters for the probability that the next edge will be added to a pendant edge, or an internal edge, then a probabilistic edge-adding model can be set up in which the Yule, uniform, and comb models are special cases. While this can readily be done, unfortunately the equations for the probabilities of the tree shapes become rather complicated, and it is unclear that the conjecture of Aldous is amenable to an approach of this sort.

8.5 Upper Bound on Probability of Fully Symmetric Trees

In this section we show that for a probability distribution with the sampling consistency property there is an upper bound on the probability of the fully symmetric tree shapes. First, as we show in the following lemma, if a probability distribution on labelled rooted trees has the group elimination property then it has the sampling consistency property. So, for example, the Yule, uniform, and comb probability models on labelled rooted trees have the sampling consistency property.

Lemma 14 *If a probability distribution on labelled rooted trees satisfies the group elimination property then it has the sampling consistency property.*

Proof. Let $k = n - 1$ then the group elimination property reduces to

$$\mathbb{P}_n[T_{|\{1,2,\ldots,n-1\}} = t \mid \{n\} \text{ is a group on T }] = \mathbb{P}_{n-1}[t].$$

A single leaf by itself always forms a group, so the leaf labelled n in particular always forms a group, thus the lemma follows. □

Before we prove the main result we need the results of the three lemmas that follow.

Lemma 15 *If a probability distribution on labelled rooted trees satisfies the sampling consistency property then*

$$\mathbb{P}_n[T_{|\{1,2,\ldots,k\}} = t] = \mathbb{P}_k[t] \,.$$

Proof. The proof is by induction on $s = n - k$. For $s = 1$ the result follows from Lemma 14. Suppose $s > 1$.

$$\begin{aligned}
\mathbb{P}_n[T_{|\{1,\ldots,k\}} = t] &= \sum_{T:T_{|\{1,\ldots,k\}}=t} \mathbb{P}_n[T] \\
&= \sum_{t':t'_{|\{1,\ldots,k\}}=t} \; \sum_{T:T_{|\{1,\ldots,k+1\}}=t'} \mathbb{P}_n[T] \\
&= \sum_{t':t'_{|\{1,\ldots,k\}}=t} \mathbb{P}_n[T_{|\{1,\ldots,k+1\}} = t'] \\
&= \mathbb{P}_{k+1}[t'_{|\{1,\ldots,k\}} = t] \\
&= \mathbb{P}_k[t] \qquad \text{(via Lemma 14).}
\end{aligned}$$

□

The proof of the lemma that follows was given in a personal communication from Rolf Kleinknecht.

Lemma 16

$$(1+u)^n(1-u)^n \leq 1 - \left[\left(\frac{1+u}{2}\right)^n - \left(\frac{1-u}{2}\right)^n\right]^2$$

where $n \geq 1$ is an integer, and $0 \leq u \leq 1$ is a real number.

Proof. For all $m > 0, x > 0$ we have $(x^{m/2} - x^{-m/2})^2 \geq 0$. From this inequality we obtain

$$x^m + x^{-m} \geq 2\,. \tag{8.12}$$

Let $0 < b \leq a < 1$, and $0 < k < n$. Putting $x = a/b$ and $m = n - k$ in (8.12) then

$$\left(\frac{b}{a}\right)^{n-k} + \left(\frac{b}{a}\right)^{-n+k} \geq 2\,.$$

Multiplying this equation by $a^n b^n$ gives

$$a^k b^{2n-k} + a^{2n-k} b^k \geq 2a^n b^n\,. \tag{8.13}$$

Leaving this inequality for a moment, assume that we also have $a + b = 1$, then

$$(a+b)^{2n} = 1\,. \tag{8.14}$$

Or, expanding by the binomial theorem instead,

$$\begin{aligned}
(a+b)^{2n} &= \sum_{k=0}^{2n} \binom{2n}{k} a^k b^{2n-k} \\
&= a^{2n} + b^{2n} + \binom{2n}{n} a^n b^n + \sum_{k=1}^{n-1} \binom{2n}{k} \left[a^{2n-k} b^k + a^k b^{2n-k}\right] \\
&\geq a^{2n} + b^{2n} + \binom{2n}{n} a^n b^n + \sum_{k=1}^{n-1} \binom{2n}{k} 2a^n b^n \qquad \text{from inequality (8.13)} \\
&= a^{2n} + b^{2n} + a^n b^n \left[\binom{2n}{n} + 2\sum_{k=1}^{n-1} \binom{2n}{k}\right] \\
&= a^{2n} + b^{2n} + a^n b^n [4^n - 2]\,.
\end{aligned} \tag{8.15}$$

Comparing (8.14) and (8.15) we get

$$a^{2n} + b^{2n} + a^n b^n (4^n - 2) \leq 1\,.$$

Putting $a = (1+u)/2$ and $b = (1-u)/2$ (which satisfy $0 < b \leq a < 1$ for $0 < u < 1$), then rearranging, we get the stated inequality of the lemma. Substituting $u = 0, 1$ in the inequality verifies that it is also true for these values. □

Lemma 17 *Let n be a positive integer. We define $g(n)$ for $n \geq w$ by*

$$g(n) = \begin{cases} 1 & n = w\,, \\ \max\limits_{\frac{w}{2} \leq i \leq n - \frac{w}{2}} \left\{ \binom{i}{w/2}\binom{n-i}{w/2} + g(i) + g(n-i) \right\} & n > w\,, \end{cases}$$

where w is a fixed integer: $w = 2^m$, $m \in \{1, 2, 3, ...\}$. If we define

$$h(n) = \frac{n^w}{[(\frac{w}{2})!]^2(2^w - 2)}$$

then $g(n) \leq h(n)$ for $n \geq w$.

Proof. The proof is by induction.

Step One. First we need to show that $g(w) = 1 \leq h(w)$. We have

$$\begin{aligned} h(w) &= \frac{w^w}{[(\frac{w}{2})!]^2(2^w - 2)} \\ &= \frac{2^w(\frac{w}{2})^w}{[(\frac{w}{2})!]^2(2^w - 2)} \\ &\geq \frac{(\frac{w}{2})^w}{[(\frac{w}{2})!]^2} \\ &\geq \left(\frac{v^v}{v!}\right)^2 \qquad \text{where } v = w/2 \geq 1 \\ &\geq 1\,. \end{aligned}$$

Step Two. We now need to show that

$$\binom{i}{w/2}\binom{n-i}{w/2} + h(i) + h(n-i) \leq h(n) \qquad \text{where } \frac{w}{2} \leq i \leq n - \frac{w}{2}\,; \quad n \geq w\,. \tag{8.16}$$

Upon substitution for the functions in h this condition becomes

$$(2^w - 2)p_w(i) + i^w + (n-i)^w - n^w \leq 0\,, \tag{8.17}$$

where $p_w(i)$ is a polynomial of degree w in i:

$$p_w(i) = \overbrace{i(i-1)\cdots(i-w/2+1)}^{\frac{w}{2}\text{ terms}}\overbrace{(i-n)(i-n+1)\cdots(i-n+w/2-1)}^{\frac{w}{2}\text{ terms}}\,.$$

Since $p_w(i) \leq i^{w/2}(n-i)^{w/2}$ then if the inequality

$$(2^w - 2)i^{w/2}(n-i)^{w/2} + i^w + (n-i)^{w/2} - n^w \leq 0 \tag{8.18}$$

is true then so is the inequality (8.17). If we let $v = w/2$, and take $i \to x$ where x is real, then another way of writing the inequality (8.18) is

$$[4x(n-x)]^v \leq n^{2v} - [x^v - (n-x)^v]^2 \quad v \geq 1 \; ; \; v \leq x \leq n-v \; . \tag{8.19}$$

To prove this inequality make the substitution $u = 2x/n - 1$, then rearrange, to obtain the equivalent inequality

$$(1+u)^v(1-u)^v \leq 1 - \left[\left(\frac{1+u}{2}\right)^v - \left(\frac{1-u}{2}\right)^v\right]^2$$

where $v \geq 1$ is an integer, $n \geq 2v$, and $-1 + 2v/n \leq u \leq 1 - 2v/n$. Note if this inequality is true for $0 \leq u \leq 1 - 2v/n$ then by symmetry it is also true for $-1 + 2v/n \leq u \leq 0$. By Lemma 16 the inequality is true for $0 \leq u \leq 1$, thus it certainly is true for $-1 + 2v/n \leq u \leq 1 - 2v/n$. Since this inequality is true then so is inequality (8.18), thereby establishing the original inequality (8.16).

Step Three. We have

$$g(n) = \binom{n}{w/2}\binom{n}{w/2} + g(n/2) + g(n/2) \quad \leq \binom{n}{w/2}\binom{n}{w/2} + h(n/2) + h(n/2) \leq h(n) \; .$$

□

Using the previous three lemmas we can now prove our main result regarding an upper bound for the probability of the fully symmetric trees.

Theorem 14 *If a probability distribution on labelled rooted trees satisfies the sampling consistency property then*

$$\mathbb{P}_w(\tau_w) \leq \frac{\binom{w}{w/2}}{2^w - 2}$$

where τ_w is the fully balanced tree shape on w leaves, and $w = 2^m$ where m is a positive integer. Asymptotically we have, as $w \to \infty$,

$$\mathbb{P}_w(\tau_w) \leq \sqrt{\frac{2}{\pi w}} \to 0 \; .$$

Proof. Suppose T is a labelled tree on n leaves and $X \subseteq \{1, 2, 3, \ldots, n\}$.

$$\begin{aligned} \text{Let } q &= \sum_{T, X:|X|=w \;\&\; T_{|X} \in \tau_w} \mathbb{P}_n[T] \\ &= \sum_{X:|X|=w} \; \sum_{T:T_{|X} \in \tau_w} \mathbb{P}_n[T] \\ &= \sum_{X:|X|=w} \mathbb{P}_w[\tau_w] \quad \text{(Lemma 15)} \\ &= \binom{n}{w} \mathbb{P}_w[\tau_w] \; . \end{aligned} \tag{8.20}$$

But we can also write

$$\begin{aligned} q &= \sum_T \mathbb{P}_n[T] \sum_{X:T_{|X} \in \tau_w} 1 \\ &\leq \max_T \left(|X| : T_{|X} \in \tau_w \right) . \end{aligned} \tag{8.21}$$

Together expressions (8.20) and (8.21) give

$$\mathbb{P}_w(\tau_w) \leq \frac{g(n)}{\binom{n}{w}} , \tag{8.22}$$

where $g(n) = \max_T(|X| : T_{|X} \in \tau_w)$.

If a tree T is split into two subtrees T_1, T_2 with number of leaves n_1, n_2 respectively then

$$g(n) = \max_i \left\{ \binom{i}{w} \binom{n-i}{w} + g(i) + g(n-i) \right\} ,$$

where $n \geq w$, $\frac{w}{2} \leq i \leq n - \frac{w}{2}$, $g(w) = 1$. So using Lemma 17 expression (8.22) becomes

$$\mathbb{P}_w(\tau_w) \leq \frac{g(n)}{\binom{n}{w}} \leq \frac{\frac{n^w}{\left[(\frac{w}{2})!\right]^2 (2^w - 2)}}{\binom{n}{w}} \qquad \text{for } n \geq w .$$

The upper limit on the right hand side is a strictly decreasing function of n, and

$$\lim_{n \to \infty} \frac{\frac{n^w}{\left[(\frac{w}{2})!\right]^2 (2^w - 2)}}{\binom{n}{w}} = \frac{\binom{w}{w/2}}{2^w - 2} ,$$

hence

$$\mathbb{P}_w(\tau_w) \leq \frac{\binom{w}{w/2}}{2^w - 2} .$$

The asymptotic result follows directly from the relationship $\binom{2n}{n}/2^{2n} \sim \frac{1}{\sqrt{\pi n}}$ with $n = \frac{w}{2}$. □

Chapter 9

A Modification of the Yule Model

9.1 Introduction

One of the assumptions of the Yule model is that the probability of speciation is the same for all lineages at any given time (see p. 11). Here we introduce a modification of the Yule model in which this assumption is not true. The motivation behind this modification is to determine what effect this has on the balance of trees, where for simplicity we look just at the effect the modification has on the probability of the symmetric tree shape on four leaves.

Taking a model in which the probability of speciation is a function of the time from the last speciation event of a lineage we introduced two variations. Firstly, we made the speciation rate constant up to a certain time, then zero afterwards. This was found to make the symmetric tree on four leaves less probable. Secondly, we made the speciation rate zero up to a certain time, then constant afterwards. This was found to make the symmetric tree more probable. In both variations the probability of the symmetric tree, given that there is a tree on four leaves, was a function of time.

9.2 The Modification

In Steel and McKenzie [68] a modification of the Yule model was introduced in which the probability of speciation was a function of the time from the last speciation event of a lineage. Starting with one lineage at time $t = 0$, let there be a probability $s(t)$ that it will split at time t resulting in two lineages. Then each of these two lineages is allowed

to independently evolve, with the probability of a split occurring equal to $s(t-t_0)$, where t_0 is the time at which the first lineage split. This process is repeated until there are n species present at time t. Note that in the case where $s(t)$ is a constant then we have the Yule model.

Let $T(t), N(t)$ be random variables whose possible values are the set of unlabelled trees at time t, and the number of unlabelled trees at time t respectively. Let τ be a particular unlabelled tree on n leaves then we define

$$f(\tau,t) = \mathbb{P}[T(t) = \tau]; \quad \nu(n,t) = \mathbb{P}[N(t) = n] .$$

To calculate the probabilities $f(\tau,t)$ and $\nu(n,t)$ first let

$$S(x) = \mathbb{P}[t_0 \geq x] = \exp[-\int_0^x s(\lambda)d\lambda]; \qquad \sigma(x) = s(x)S(x) .$$

Then for the tree τ with the subtrees τ_1, τ_2 we have the recursions [68]

$$f(\tau,t) = 2^{\delta(\tau)} \int_0^t f(\tau_1, t-x) f(\tau_2, t-x)\sigma(x)S(x)dx , \tag{9.1}$$

$$\nu(n,t) = \sum_{i=1}^{n-1} \int_0^t \nu(i, t-x)\nu(n-i, t-x)\sigma(x)dx , \tag{9.2}$$

where

$$\delta(\tau) = \begin{cases} 1 & \text{if } \tau_1 \neq \tau_2 , \\ 0 & \text{otherwise.} \end{cases}$$

We also define

$$p_n(\tau,t) = \mathbb{P}[T(t) = \tau \mid T(t) \text{ has } n \text{ leaves}] = \frac{f(\tau,t)}{\nu(n,t)} . \tag{9.3}$$

9.3 Explosive Radiation

One functional form for $s(t)$ investigated in [68] involved a refractory period such that if a lineage had not speciated up to the time ϵ then it never would:

$$s(t) = \begin{cases} s, & t \leq \epsilon \\ 0, & t > \epsilon \, . \end{cases} \tag{9.4}$$

In the limiting case as $\epsilon \longrightarrow \infty$ we get the Yule model. Such a functional form could model the scenario in which an organism that has recently speciated (say by the appearance of a new ecological niche), is more likely to speciate then an organism that has not speciated for a long time.

It was found in [68] that if the condition $t > n\epsilon$ was met, where n is the number of species, then the conditional probability distribution for the trees (i.e. $p_n(\tau, t)$) was in fact the uniform distribution. So making the speciation rate zero for some finite time after a speciation event gives, for t large enough, trees that are less balanced (since trees in the uniform model are on average less balanced then those in the Yule model). Here we investigate the probability distribution, both conditional and unconditional, of the symmetric tree on four leaves for all $t > 0$.

Let τ_4 be the symmetric tree on four leaves (with a 'ghost' edge representing the initial species present at $t = 0$). From the recursion (9.1) we obtain, with the aid of the computer algebra package *MAPLE*, the following expression for the probability of the symmetric tree on four leaves as a function of time:

$$f(\tau_4, t) = \begin{cases} \frac{1}{3}\, e^{-st} - e^{-2\,st} + e^{-3\,st} - \frac{1}{3}\, e^{-4\,st} & t \leq \epsilon \, , \\ \frac{8}{3}\, e^{-s(3\,\epsilon+t)} - 4\, e^{-s(3\,\epsilon+t)} st + 4\, e^{-s(3\,\epsilon+t)} s\epsilon - \frac{2}{3}\, e^{-s(t+\epsilon)} & \\ -4\, e^{-2\,s(t+\epsilon)} + e^{-4\,s\epsilon} + 2\, e^{-3\,st} - \frac{1}{3}\, e^{-2\,s(2\,t-\epsilon)} - 2\, e^{-s(2\,t+\epsilon)} & \\ +e^{-s(2\,t-\epsilon)} - e^{-s(-2\,\epsilon+3\,t)} + \frac{1}{3}\, e^{-s(-3\,\epsilon+4\,t)} + e^{-s(2\,\epsilon+t)} & \epsilon < t \leq 2\epsilon \, , \\ -\frac{10}{3}\, e^{-s(4\,\epsilon+t)} + e^{-6\,s\epsilon} - 3\, e^{-5\,s\epsilon} + e^{-4\,s\epsilon} + 4\, e^{-s(2\,t+\epsilon)} & \\ -12\, e^{-s(3\,\epsilon+t)} s\epsilon - 2\, e^{-s(-2\,\epsilon+3\,t)} + 2\, e^{-2\,st} + \frac{1}{3}\, e^{-s(-5\,\epsilon+4\,t)} & \\ +4\, e^{-s(3\,\epsilon+t)} st & 2\epsilon < t \leq 3\epsilon \, , \\ \left(e^{-s\epsilon}\right)^4 \left(1 - e^{-s\epsilon}\right)^3 & t > 3\epsilon \, . \end{cases} \tag{9.5}$$

Plotting $f(\tau_4, t)$ for $\epsilon = 0.2, 0.5, 1.0$ we obtain Figure 9.1. From the figure it can be seen that as ϵ is decreased from infinity (the Yule model value) the maxima decreases in

size and moves to the left. For $t > 3\epsilon$ the curve is constant (see equation (9.5)) and equal to

$$f^{\infty}(k_1) = \left(e^{-k_1}\right)^4 \left(1 - e^{-k_1}\right)^3, \qquad \text{where } k_1 = s\epsilon.$$

Plotting $f^{\infty}(k_1)$ gives Figure 9.2. The expression for the probability of the symmetric tree on four leaves, conditional on there being a tree on four leaves at time t ($p_n(\tau_4, t)$), is rather complicated. Because of this we feel little is to be gained by displaying it here. However the functional form can readily be plotted which we do in Figure 9.3.

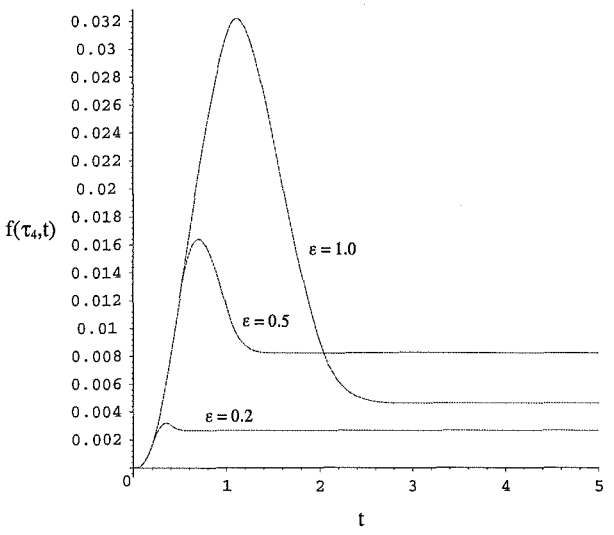


Figure 9.1: Explosive radiation and the probability of the symmetric tree on four leaves as a function of time. The refractory period is of length ϵ, and the speciation rate (s) is equal to one. The probability is constant for $t > 3\epsilon$ (see Figure 9.2).

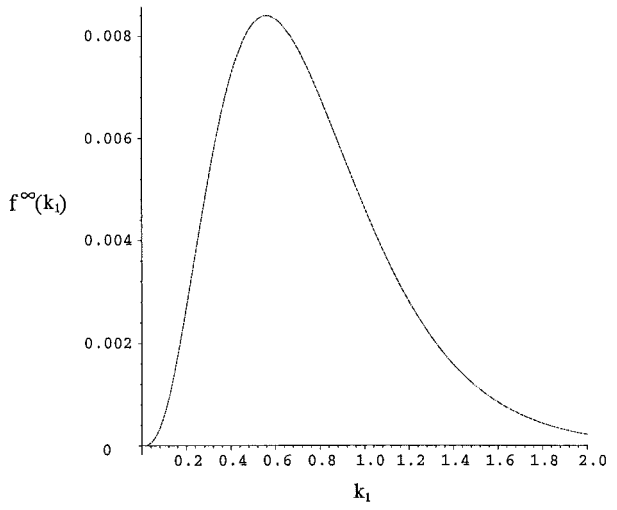


Figure 9.2: Explosive radiation and the probability of the symmetric tree on four leaves for $t > 3\epsilon$. The probability depends only on $k_1 = s\epsilon$. The maxima of $6912/823543 \approx 0.0084$ occurs at $k_1 = \ln(7/4)$.

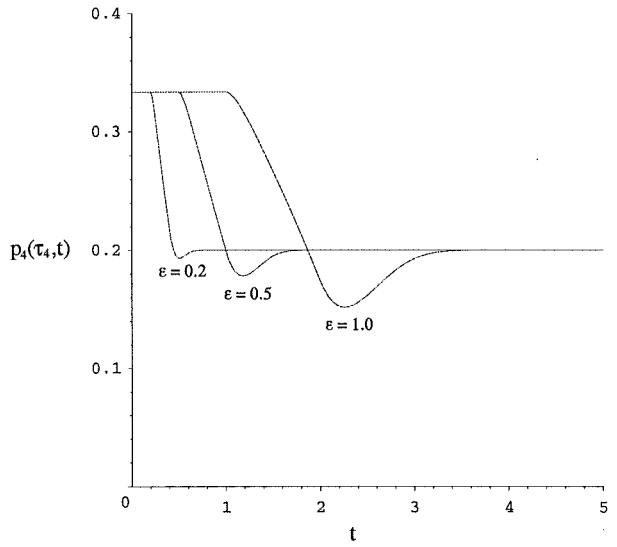


Figure 9.3: Explosive radiation and the the probability of the symmetric tree on four leaves, conditional on there been a tree on four leaves at time t. The refractory period is ϵ, and the speciation rate (s) is equal to one. The probability is $1/3$ for $t \leq \epsilon$, and $1/5$ for $t > 4\epsilon$.

9.4 Delayed Speciation

Here we investigate the effect of having a refractory period after a speciation event up to the time β, then a constant speciation rate afterwards. Such a form for the refractory period has being investigated before by simulation, but not by analytical methods [44].

The functional form for the speciation rate in this case is

$$s(t) = \begin{cases} 0, & t \leq \beta \\ s, & t > \beta \, . \end{cases} \tag{9.6}$$

For $\beta = 0$ we have the Yule model. Such a functional form might model what happens in peripatric speciation in which isolated species on the periphery of a population undergo speciation events [28, 44]. In such a form of speciation recently formed species need time to stabilise their genetic organisation, and will only speciate again once their geographic spread is large enough for further isolated populations to form.

From the recursion (9.1) we obtain, using the computer algebra package *MAPLE* again, the following expression for the probability of the symmetric tree on four leaves as a function of time:

$$f(\tau_4, t) = \begin{cases} 0 & t \leq 2\beta \, , \\ e^{s\beta}\left(-2\,e^{s(-t+\beta)}st + 4\,e^{s(-t+\beta)}s\beta + e^{-s\beta} - e^{s(-2\,t+3\,\beta)}\right) & 2\beta < t \leq 3\beta \, , \\ \frac{10}{3}\,e^{s(-t+3\,\beta)} - 3\,e^{s(-t+2\,\beta)} - 4\,e^{2\,s(-t+3\,\beta)} & \\ +4\,e^{s(5\,\beta-2\,t)} + 2\,e^{3\,s(-t+3\,\beta)} - e^{2\,s(-t+2\,\beta)} - e^{s(8\,\beta-3\,t)} & \\ -\frac{1}{3}\,e^{4\,s(-t+3\,\beta)} - 2\,e^{s(-t+2\,\beta)}s\beta \, . & t > 3\beta \, . \end{cases}$$

As in the previous section the expression for the probability of the symmetric tree on four leaves, conditional on there being a tree on four leaves at time t ($p_n(\tau_4, t)$), is rather complicated. However, the asymptotic (in time t) form for this expression is much simpler and we have

$$p_\infty(k) = \lim_{t\to\infty} p_n(\tau_4, t) = \frac{e^{k_2}[10e^{k_2} - 9 - 6k_2]}{10e^{2k_2} - 3e^{k_2} - 6k_2e^{k_2} - 4}, \qquad \text{where } k_2 = s\beta.$$

Plotting $f(\tau_4, t)$ for $\beta = 0.0, 0.2, 0.4$ gave Figure 9.4. Not surprisingly, as β is increased the probability of the symmetric tree increases at latter times. However, instead of the

probability curve for $\beta = 0$ simply being shifted to the right, the size of the central maxima increases as well.

Plotting $p_n(\tau_4, t)$, again for $\beta = 0.0, 0.2, 0.4$, gave Figure 9.5. When β is non-zero the conditional probability is a function of time. This sounds a wary note for simulations which estimate conditional probabilities based on an ensemble of trees shapes generated at a variety of times - this gives the average conditional probability for a given value of β [44].

The asymptotic values ($p_\infty(k)$) for the conditional probabilities are shown in Figure 9.6. Note that the asymptotic value depends on the product $s\beta$, so increasing the speciation rate, with a fixed refractory period, also leads to increased symmetry. It was found in [60] that for large refractory periods the symmetry *decreased* for trees, seemingly in contradiction to the results we have obtained. However, in [60] only one of the two descendant species underwent a refractory period, where as here both of the descendant species undergo refractory periods in which further speciation can not occur.

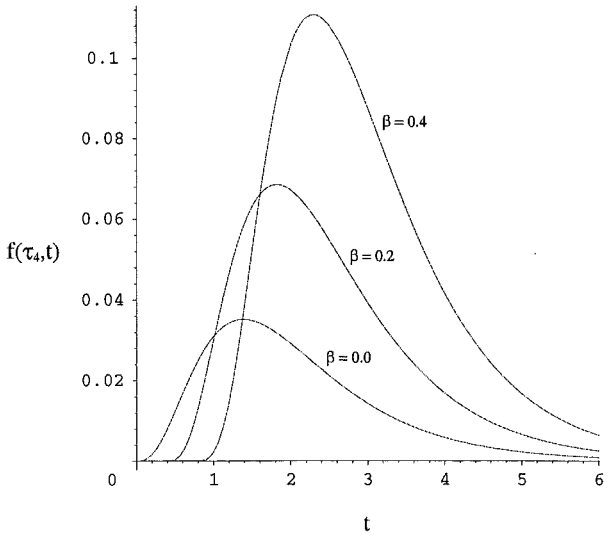


Figure 9.4: Delayed speciation and the probability of the symmetric tree on four leaves as a function of time. The refractory period is of length β, and the speciation rate (s) is equal to one. The probability is zero for $t \leq 2\beta$.

We have shown that for the tree shapes on four leaves increasing the size of the refractory period at the start (β) leads to the symmetric tree becoming more probable. Although we have not investigated the effect of increasing β on tree shapes on more than four leaves,

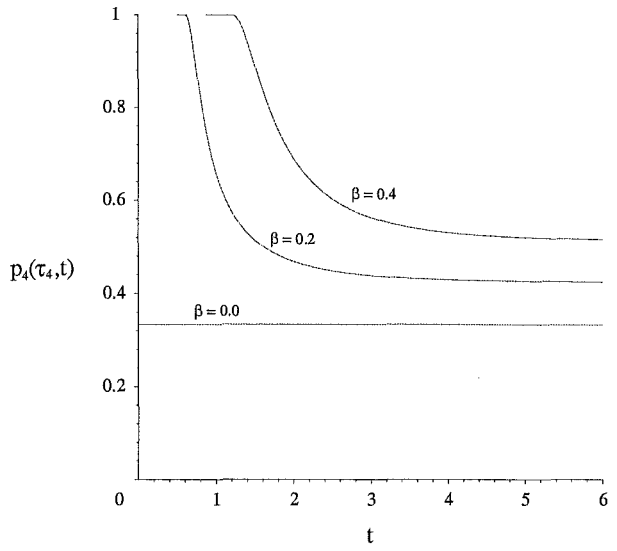


Figure 9.5: Delayed speciation and the probability of the symmetric tree on four leaves, conditional on there been a tree on four leaves at time t. The refractory period is β, and the speciation rate (s) is equal to one. The probability is zero for $t \leq 2\beta$, and equal to one for $2\beta < t \leq 3\beta$. For the asymptotic value of the probability as $t \to \infty$ see Figure 9.6.

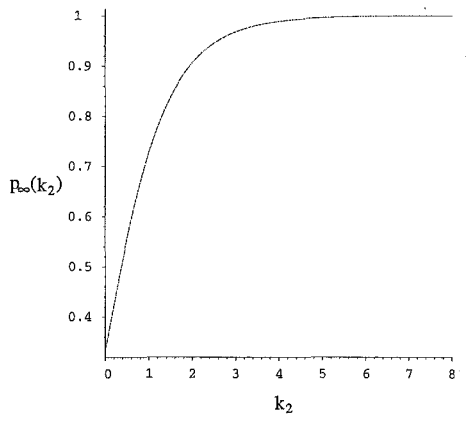


Figure 9.6: Delayed speciation and the asymptotic (in time t) probability for the symmetric tree on four leaves, conditional on there being a tree having four leaves. The asymptotic probability only depends on the value of k_2, where $k_2 = s\beta$.

we speculate that the probability of the more symmetric shapes will increase. This seems likely since the most symmetric trees on n leaves largely grow from the most symmetric trees on $n-1$ leaves. Furthermore, another way of characterising the effect of increasing β is to say that this increases the probability of retaining cherries (see Section 1.3.1); but the trees with the most number of cherries are the symmetric trees.

9.5 Discussion

In summary: (1) A constant speciation rate followed by a refractory period leads to *more* imbalanced trees (2) An initial refractory period followed by a constant rate of speciation leads to *less* imbalanced trees. Also, whether or not both descendant species undergo a refractory period, appears to have an important effect on tree balance when the refractory period is longer. Thus introducing a refractory period into the Yule model can have a variety of effects, depending on just what form it takes.

An obvious question is what happens to tree balance when a refractory period is included before and after a period of non-zero speciation probability? It would seem the answer would depend on the relative length of the refractory periods, with the possibility that tree balance would be unaffected if the relative lengths were set correctly. The answer to this question, and the possible biological interpretations, awaits further work.

Bibliography

[1] A. Aho, T. Sagiv, T. Szymanski, and J. Ullman. Inferring a tree from lowest common ancestors with an application to the optimization of relational expressions. *SIAM Journal of Computing*, 10(3):405–421, 1981.

[2] D. Aldous. The continuum random tree II: an overview. In M. T. Barlow and N. H. Bingham, editors, *Stochastic Analysis*, pages 23–70. Cambridge University Press, Cambridge, 1991.

[3] D. Aldous. The continuum random tree III. *Ann. Probab.*, 21:248–289, 1993.

[4] D. Aldous. Probability distributions on cladograms. In D. Aldous and R. Permantle, editors, *Random Structures*, volume 76, pages 1–18. Springer, 1996.

[5] A. Amir and D. Keselman. Maximum agreement subtree in a set of evolutionary trees: metrics and efficient algorithms. *Siam J. Comput.*, 26(6):1656–1669, 1997.

[6] D. Applebaum. *Probability and Information: an integrated approach*, chapter 6. Cambridge University Press, Cambridge; New York, 1996.

[7] K. B. Athreya and S. Karlin. Embedding of urn schemes into continuous time Markov branching processes and related limit theorems. *Ann. Math. Statist.*, 39:1801–1817, 1968.

[8] K. B. Athreya and P. E. Ney. *Branching Processes*, pages 219–224. Springer–Verlang, 1972.

[9] A. Bagchi and A. K. Pal. Asymptotic normality in the generalized Polya-Eggenburger urn model, with an application to computer data structures. *SIAM J. Algebraic Discrete Methods*, 6:394–405, 1985.

[10] Z. Bai and F. Hu. Asymptotic theorems for urn models with nonhomogeneous generating matrices. *Stochastic Processes and their Applications*, 80:87–101, 1999.

[11] H. Bodlaender, M. Fellows, and T. Warnow. Two strikes against perfect phylogeny. In W. Kuich, editor, *Automata,languages and Programming: 19th international colloquium*, number 623 in Lecture Notes in Computer Science, pages 273–287, Berlin ; New York, 1992. Springer-Verlag.

[12] D. R. Brooks and E. O. Wiley. *Evolution as Entropy: Toward a unified theory of biology.* The University of Chicago Press, Chicago, 1988.

[13] J. K. M. Brown. Probabilities of evolutionary trees. *Syst. Biol.*, 43(1):78–90, 1994.

[14] L. Cavalli-Sforza and A. Edwards. Phylogenetic analysis. *Evolution*, 21:550–570, 1967.

[15] Y. S. Chow and H. Teicher. *Probability Theory: Independence, Interchangeability, Martingales.* Springer-Verlag, New York, 1988.

[16] D. H. Colless. Phylogenetics: The theory and practice of phylogenetic systematics II (book review). *Syst. Zool.*, 31:100–104, 1982.

[17] M. Constantinescu and D. Sankoff. Tree enumeration modulo a consensus. *Journal of Classification*, 3:349–356, 1986.

[18] M. Constantinescu and D. Sankoff. An efficient algorithm for supertrees. *Journal of Classification*, 12(1):101–112, 1995.

[19] T. Cover and J. Thomas. *Elements of information theory*, chapter 1, pages 1–11. John Wiley and Sons, New York, 1991.

[20] A. Edwards and L. Cavalli-Sforza. Reconstruction of evolutionary trees. In W. Heywood and J. McNeill, editors, *Phenetic and phylogenetic classification*, number 6, pages 67–76. 1964.

[21] A. W. F. Edwards. Estimation of the branch points of a branching diffusion process. *J.R. Stat. Soc. Ser. B*, 32:155–174, 1970.

[22] R. B. Eggleton and R. K. Guy. Catalan strikes again! How likely is a function to be convex? *Math. Mag.*, 61:211–219, 1988.

[23] J. S. Farris. Estimating phylogenetic trees from distance matrices. *American Naturalist*, 106:646–668, 1972.

[24] W. Feller. *An Introduction to Probability Theory and its Applications*, chapter 11, page 54. John Wiley and Sons, Inc, 1968.

[25] C. R. Finden and A. D. Gordon. Obtaining common pruned trees. *Journal of Classification*, 2:255–276, 1985.

[26] M. Fisz. *Probability theory and mathematical statistics*, chapter 6, page 206. John Wiley and Sons, Inc, New York, 1963.

[27] J. E. Freund and R. E. Walpole. *Mathematical Statistics*, chapter 4, page 150. Prentice-Hall, Englewood Cliff, N.J., 4 edition, 1987.

[28] D. J. Futuyma. *Evolutionary Biology*, chapter 16, pages 493–496. Sinauer Assoicates, Inc, Sunderland, Massachusetts, 1997.

[29] W. Goddard and K. Grzegorz. The minimum size of agreement subtrees of two binary trees. In *Proceedings of the Twenty-fourth Southeastern International Conference on Combinatorics, Graph Theory, and Computing*, volume 97, pages 131–136, Boca Raton, FL, 1993.

[30] S. J. Gould, D. M. Raup, J. J. Sepkowski, T. J. M. Schopf, and D. S. Simberloff. The shape of evolution: a comparison of real and random clades. *Paleobiology*, 3:23–40, 1977.

[31] R. L. Graham, D. E. Knuth, and O. Patashnik. *Concrete Mathematics*. Addison-Wesley, 1989.

[32] G. Grimmett and D. Stirzaker. *Probability and Random Processes*. Claredon Press, Oxford, 1982.

[33] C. Guyer and J. B. Slowinski. Comparisons of observed phylogenetic topologies with null expectations among three monophyletic lineages. *Evolution*, 45(2):340–350, 1991.

[34] J. Haigh. The recovery of the root of a tree. *J. Appl. Prob.*, 7:79–88, 1970.

[35] P. Hall and C. C. Heyde. *Martingale Limit Theory and its applications*, chapter 3. Academic Press, New York, 1980.

[36] E. F. Harding. The probabilities of rooted tree-shapes generated by random bifurcation. *Adv. Appl. Prob.*, 3:44–77, 1971.

[37] M. Härlin. Biogeographic patterns and the evolution of eureptanic nemerteans. *Biological Journal of the Linnean Society*, 58:325–342, 1996.

[38] S. B. Heard. Patterns in tree balance among cladistic, phenetic, and randomly generated phylogenetic trees. *Evolution*, 46(6):1818–1826, 1992.

[39] S. B. Heard. Patterns in phylogenetic tree balance with variable and evolving speciation rates. *Evolution*, 50(6):2141–2148, 1996.

[40] M. D. Hendy and D. Penny. Branch and bound algorithms to determine minimal evolutionary trees. *Math. Biosci.*, 59:277–290, 1982.

[41] J. P. Huelsenbeck and M. Kirkpatrick. Do phylogenetic methods produce trees with biased shapes? *Evolution*, 50(4):1418–1424, 1996.

[42] J. Kingman. On the genealogy of large populations. *J. Appl. Prob.*, 19A:27–43, 1982.

[43] M. Kirkpatrick and M. Slatkin. Searching for evolutionary patterns in the shape of a phylogenetic tree. *Evolution*, 47(4):1171–1181, 1993.

[44] J. B. Losos and F. R. Adler. Stumped by trees? A generalized null model for patterns of organismal diversity. *The American Naturalist*, 145(3):329–342, Mar 1995.

[45] W. Lynch. More combinatorial properties of certain trees. *The Computer Journal*, 71:299–302, 1965.

[46] A. McKenzie and M. Steel. Distributions of cherries for two models of trees. *Mathematical Biosciences*, 164(1):81–92, 2000.

[47] A. Ø. Mooers. Tree balance and tree completeness. *Evolution*, 49(2):379–384, 1995.

[48] A. Ø. Mooers and S. B. Heard. Inferring evolutionary process from phylogenetic tree shape. *The Quarterly Review of Biology*, 72(1):31–54, Mar 1997.

[49] S. Nee, R. M. May, and P. H. Harvey. The reconstructed evolutionary process. *Phil. Trans. R. Soc. Lond. B*, 344:305–311, 1994.

[50] M. Ng and N. Wormald. Reconstruction of rooted trees from subtrees. *Discrete Applied Mathematics*, 69(1–2):19–31, 1996.

[51] A. M. Odlyzko. Asymptotic enumeration methods. In R. L. Graham, M. Grötschel, and L. Lovász, editors, *Handbook of Combinatorics: Volume II.* The MIT Press, Cambridge, Massachusetts, 1995.

[52] R. D. M. Page. Random denrograms and null hypotheses in cladistic biogeography. *Syst. Zool.*, 40(1):54–62, 1991.

[53] R. D. M. Page and E. C. Holmes. *Molecular Evolution: a phylogenetic approach*, chapter 2, pages 11–36. Blackwell Science, Oxford ; Malden, MA, 1998.

[54] Y. Van de Peer and R. De Wacher. Treecon: a software package for the construction and drawing of evolutionary trees. *Comput. Applic. Biosci.*, 9:177–182, 1993.

[55] E. C. Pielou. *Mathematical Ecology*, chapter 19, pages 292–299. John Wiley and Sons, New York, 1977.

[56] D. M. Raup, S. J. Gould, T. J. M. Schopf, and D. S. Simberloff. Stochastic models of phylogeny and the evolution of diversity. *The Journal of Geology*, 81(5):525–542, Sep 1973.

[57] M. Ridley. *Evolution*, chapter 17, pages 460–466. Blackwell Science,Inc., Cambridge, Massachusetts, USA, 1993.

[58] J. S. Robertson, K. Bolinger, L. M. Glasser, and N. J. A. Sloane. Discrete mathematics. In D. Zwillinger, editor, *CRC Standard mathematical tables and formulae.* Boca Raton: CRC Press, New York, 30th edition, 1996.

[59] J. S. Rogers. Central moments and probability distribution of Colless's coefficient of tree imbalance. *Evolution*, 48(6):2026–2036, 1994.

[60] J. S. Rogers. Central moments and probability distributions of three measures of phylogentic tree imbalance. *Systematic Biology*, 45:99–110, 1996.

[61] H. M. Savage. The shape of evolution: systematic tree topology. *Biological Journal of the Linnean Society*, 20:225–244, 1983.

[62] K. Shao and R. R. Sokal. Tree balance. *Syst. Zool.*, 39:266–276, 1990.

[63] J. B. Slowinski. Probabilities of n-trees under two models: a demonstration that asymmetrical interior nodes are not improbable. *Syst. Zool.*, 39(1):89–94, 1990.

[64] J. B. Slowinski and C. Guyer. Testing the stochasticity of patterns of organismal diversity: an improved null model. *The American Naturalist*, 134(6):907–921, Dec 1989.

[65] A. Smith. Rooting molecular trees: problems and strategies. *Biological Journal of the Linnean Society*, 51:279–292, 1994.

[66] R. T. Smythe. Central limit theorems for urn models. *Stochastic Processes and their Applications*, 65:115–137, 1996.

[67] M. Steel. The complexity of reconstructing trees from qualitative characters and subtrees. *Journal of Classification*, (9):91–116, 1992.

[68] M. Steel and A. McKenzie. Properties of phylogenetic trees generated by Yule-type speciation models. *Mathematical Biosciences*, 170(1):91–112, 2001.

[69] M. Steel and D. Penny. Distribution of tree comparison metrics - some new results. *Syst. Biol.*, 42(2):126–141, 1993.

[70] M. Steel and T. Warnow. Kaikoura tree theorems: computing the maximum agreement subtree. *Information Processing Letters*, 48:77–82, 1993.

[71] M. A. Steel. Distribution of the symmetric difference metric on phylogenetic trees. *SIAM J. Discr. Math.*, 1(4):541–551, 1988.

[72] D. L. Swofford, G. J. Olsen, P. J. Waddell, and D. M. Hillis. Phylogenetic inference. In D. M. Hillis, C. Moritz, and B. K. Mable, editors, *Molecular Systmatics*, page 488. Sinauer Associates, Inc., Sunderland, Massachusetts, U.S.A, 1996.

[73] F. Tajima. Evolutionary relationships of DNA sequences in finite populations. *Genetics*, 105:437–460, 1983.

[74] G. K. Tzanetopoulos, J. Goldberg, J. J. Rushanan, and M. Hausner. Discrete mathematics. In D. Zwillinger, editor, *CRC Standard mathematical tables and formulae*, pages 161–247. Boca Raton: CRC Press, New York, 30th edition, 1996.

[75] C. O. Webb. Exploring the phylogenetic structure of ecological communities: an example for rain forest trees. *The American Naturalist*, 156(2):145–155, August 2000.

[76] J. H. M. Wedderburn. The functional equation $g(x^2) = 2x + [g(x)]^2$. *Ann. Math. 2nd ser*, 24:121–140, 1922.

[77] H. Yockey. *Information Theory and Molecular Biology.* Cambridge University Press, Cambridge, 1992.

[78] W. van Zwet. A proof of Kakutani's conjecture on random subdivision of longest intervals. *The Annals of Probability*, 6(1):133–137, 1978.

Appendix A

Distance Relationships

In this appendix we collect together some tables for the distance relationships involving the Yule and uniform models. These enable one to get a feel for how the distances vary with the number of leaves, and are designed as a reference source for those who require some actual numbers instead of formulae.

A.1 Mean Distance From the Root

Number of leaves (n)	μ_n	$2\ln n - 2(1-\gamma)$	σ_n^2	$2\ln n + 2(1+\gamma) - \frac{2\pi^2}{3}$
3	1.67	1.35	0	0
4	2.17	1.93	0.47	-0.65
5	2.57	2.37	0.71	-0.21
10	3.86	3.76	1.66	1.18
50	7.00	6.99	4.50	4.40
100	8.37	8.36	5.83	5.79

Table A.1: Mean distance and variance of a randomly chosen leaf from the root. Exact and asymptotic results for the Yule model.

Number of leaves (n)	ν_n	$\sqrt{\pi n} - \frac{3}{8}\sqrt{\frac{\pi}{n}} - 1$
3	1.67	1.69
4	2.20	2.21
5	2.66	2.67
10	4.39	4.39
50	11.44	11.44
100	16.66	16.66

Table A.2: Mean distance of a randomly chosen leaf from the root. Exact and asymptotic results for the uniform model.

A.2 Distance Between Two Leaves

Number of leaves (n)	d_n	$\mathtt{d_n}$
3	2.67	2.67
4	3.22	3.20
5	3.70	3.66
10	5.43	5.39
50	10.57	12.44
100	13.09	17.66

Table A.3: Mean distance between two (different) randomly chosen leaves under the Yule model (d_n) and uniform model ($\mathtt{d_n}$.)

Appendix B

Sum of Squared Probabilities: Yule Model

A recursion for the sum of squared probabilities for the Yule model was calculated in Section 6.2.2. Here we compute the numbers given by the recursion.

Number of leaves (n)	$\sum_t P_n[t]^2$
1	1
2	1
3	1/3
4	2/27
5	13/1080
6	7/4500
7	851/5103000
8	1.53×10^{-5}
9	1.22×10^{-6}
10	8.70×10^{-8}
15	3.84×10^{-14}
20	2.90×10^{-21}
50	7.74×10^{-73}

Table B.1: Sum of squared labelled tree probabilities for the Yule model on rooted trees.

Appendix C

Direct Proof of Asymptotic Cherry Distribution

In this section we give some of the details of a direct proof, using martingale differences, that the number of cherries in the Yule and uniform models follow a normal distribution in the asymptotic limit. As we were later able to draw upon the urn model theorems in [66], which were themselves based upon the use of martingales, we did not pursue the details of this proof further. However, in other situations where the conditions for the urn model theorems do not apply, a more direct approach like we have outlined here would be an avenue to explore.

First we explain what a martingale difference is and give a martingale difference for the numbers of cherries on a tree (Section C.1). We then state a martingale difference central limit theorem and the three conditions under which it applies (Section C.2). Lastly, we prove that the first two of these conditions do indeed apply for the martingale difference we have defined (Section C.3).

C.1 Martingales Differences

The origin of martingales can be traced, like many topics in probability, to an interest in gambling games. In particular it is associated with a type of betting scheme in which the size of the bet is doubled after each loss, where the chance of a loss at each bet of the game is 1/2. If we let X_n denote the amount of money the player has after the nth play of this game then the sequence $\{X_n : n \geq 1\}$ is a martingale. Note the es-

sential point that, since the chances of winning or loss at each bet are both 1/2, then $\mathbb{E}[X_n \mid X_1, X_2, \ldots, X_{n-1}] = X_{n-1}$. Generalising from these historical origins, we have the following contemporary definition of a martingale [32, p. 200].

Let $\{F_n\}$ be a sequence of random variables. Let X_n be some function ϕ_n of these, that is, $X_n = \phi_n(F_1, \ldots, F_n)$. The sequence $\{X_n : n \geq 1\}$ is a *martingale* with respect to the sequence $\{F_n : n \geq 1\}$ if, for all $n \geq 1$,

(a) $\mathbb{E}[|X_n|] < \infty$

(b) $\mathbb{E}[X_{n+1} \mid F_1, F_2, \ldots, F_n] = X_n$.

What we are more concerned with is a *martingale difference* [15, p. 318] which is defined as a random variable sequence such that

$$\mathbb{E}[X_{n+1} \mid F_1, F_2, \ldots, F_n] = 0 \, .$$

The connection with martingales is that if w_n is a martingale sequence, then $X_{n+1} = w_{n+1} - w_n$ can easily be shown to be a martingale difference.

C.2 Central Limit Theorem

Before we state a central limit theorem for martingale differences we need to define what it means for a sequence of discrete random variables $\{X_n\}$ to satisfy the *Lindeberg* condition [15, p. 295],[26]. Let $\{X_n, n \geq 1\}$ be a set of discrete random variables, where the probability that the nth random variable takes the value x_{nl} is p_{nl}. Also let $\mathbb{E}[X_n] = 0$, $\mathbb{E}[X_n^2] = \sigma_n^2 < \infty$. The set $\{X_n\}$ is said to obey the Lindeberg condition if

(i) $s_n^2 = \sum_{i=1}^{n} \sigma_j^2 > 0$ for some n,

(ii) $\lim_{n\to\infty} \frac{1}{s_n^2} \sum_{j=1}^{n} \sum_{x_{jl}:|x_{jl}|>\epsilon\, s_n} x_{jl}^2 p_{jl} = 0 \qquad \forall \epsilon > 0 \, .$

There are many different types of central limit theorems for martingales and martingale differences [35]. Here we state a central limit theorem which is the most useful one for our purposes [15, p. 318].

Theorem 15 *Let* $\{X_n\}$ *be a sequence of random variables. Set* $\sigma_n^2 = \mathbb{E}[X_n^2]$, $s_n^2 = \sum_{j=1}^n \sigma_j^2$. *Let* $\mathcal{N}(\mu, \sigma^2)$ *denote a normal distribution with a mean of* μ *and a variance of* σ^2. *If we have*

(i) $\mathbb{E}[X_{n+1}|F_1, F_2, \ldots, F_n] = 0,\ n \geq 1,$

(ii) $\{X_n\}$ *obey the Lindeberg condition,*

(iii) $\lim_{n\to\infty} \frac{1}{s_n^2} \sum_{j=1}^n \mathbb{E}[\,|\mathbb{E}[X_j^2 \mid F_1, F_2, \ldots, F_{j-1}]|\,] = 0,$

then $(1/s_n)\sum_{j=1}^n X_j \to \mathcal{N}(0,1)$.

C.3 Some Details of the Proof

We now prove that conditions (i) and (ii) are satisfied for a martingale difference we define involving the number of cherries in the Yule and uniform models. We introduce the random variable X_n:

$$X_n = C_n - C_{n-1} - 1 + \frac{2C_{n-1} + cn + d}{an + b}$$

where a, b, c, d are real constants. For the Yule model we set $a = 1, b = -1, c = 0, d = 0$ and for the uniform model we set $a = 2, b = -5, c = 1, d = -4$. The reasons for these choices of values for a, b, c, d will become clear in the next section, where we show that X_n is a martingale difference. We also show that the X_n satisfy the Lindeberg condition.

C.3.1 A Martingale Difference

Before we show that X_n is a martingale difference with respect to $\{C_1, C_2, \ldots, C_{n-1}\}$, we give an interpretation of the last two terms that make up X_n. Let p_n be the probability that the number of cherries for a tree on n leaves is one more than the number of cherries for the tree on $n-1$, given that the tree on $n-1$ leaves has C_{n-1} cherries. That is,

$$p_n = \mathbb{P}[C_n - C_{n-1} = 1 \mid C_1, C_2, \ldots, C_{n-1}]\,.$$

For the Yule and uniform model p_n can be explicitly calculated, as we show in the following lemma.

Lemma 18 *For the rooted Yule model and the unrooted uniform model we have*

$$p_n = 1 - \frac{2C_{n-1} + cn + d}{an + b},$$

where for the Yule model $a = 1, b = -1, c = 0, d = 0$ *and for the uniform model we have* $a = 2, b = -5, c = 1, d = -4$.

Proof. In this proof we are viewing the Yule and uniform models as processes of edge addition. For the Yule model, a new cherry only forms if the next edge is added to a non-pendant edge. If the tree on $n-1$ leaves has C_{n-1} cherries then the probability of this is $1 - 2C_{n-1}/(n-1)$. For the uniform model, a new cherry *does not* form if the next edge is added to an internal edge or to a cherry. If a tree on $n-1$ leaves has C_{n-1} cherries then the probability of this is $(2C_{n-1} + n - 4)/(2n-5)$, and taking this away from one gives p_n. □

We now see that we can write $X_n = C_n - C_{n-1} - p_n$. From this it is easy to show that X_n is a martingale difference, with respect to $\{C_1, C_2, \ldots, C_n\}$, since

$$\begin{aligned}
\mathbb{E}[X_n \mid C_1, C_2, \ldots, C_{n-1}] &= \mathbb{E}[C_n | C_1, C_2, \ldots C_{n-1}] - C_{n-1} - p_n \\
&= C_{n-1}(1 - p_n) + (C_{n-1} + 1)p_n - C_{n-1} - p_n \\
&= 0 .
\end{aligned}$$

C.3.2 Lindeberg Condition

To prove that $\{X_n\}$ satisfy the Lindeberg condition we first need to show that X_n is uniformly bounded, and that asymptotically $\mathbb{E}[X_n^2]$ is constant. We do so in the two lemmas that follow.

Lemma 19 *The random variable*

$$X_n = C_n - C_{n-1} - 1 + \frac{2C_{n-1} + cn + d}{an + b},$$

where a, b, c, d *are real constants, is uniformly bounded.*

Proof. The random variable X_n is uniformly bounded if $X_n \leq \gamma$ for some constant γ and for all n. Taking absolute values we get

$$|X_n| \leq |C_n - C_{n-1} - 1| + \left| \frac{2C_{n-1} + cn + d}{an + b} \right|. \tag{C.1}$$

Firstly, $|C_n - C_{n-1} - 1| \leq 1$ as $C_n - C_{n-1}$ equals either 0 or 1. The second term can be bounded as follows

$$\begin{aligned} \left|\frac{2C_{n-1} + cn + d}{an + b}\right| &= \left|\frac{2C_{n-1}}{an + b} + \frac{cn + d}{an + b}\right| \\ &\leq \left|\frac{2C_{n-1}}{an + b}\right| + \left|\frac{cn + d}{an + b}\right| . \end{aligned} \tag{C.2}$$

Now for the first of these bounding terms we have

$$\left|\frac{2C_{n-1}}{an + b}\right| \leq \frac{2C_{n-1}}{|a|n + |b|} \leq \frac{n - 1}{|a|n + |b|} \leq \frac{n}{|a|n + |b|} = \frac{1}{|a| + \frac{|b|}{n}} \leq \frac{1}{|a|} ,$$

and for the second we have

$$\left|\frac{cn + d}{an + b}\right| \leq \frac{|c|n + |d|}{|a|n + |b|} = \frac{|c| + \frac{|d|}{n}}{|a| + \frac{|b|}{n}} \leq \frac{|c| + |d|}{|a|} .$$

So putting back in both of these bounding terms in (C.2) we get

$$\left|\frac{2C_{n-1} + cn + d}{an + b}\right| \leq \frac{|c| + |d| + 1}{|a|} .$$

Thus substituting back into (C.1) we get

$$|X_n| \leq 1 + \frac{|c| + |d| + 1}{|a|} \qquad \forall n .$$

□

We know from the explicit formulae in Chapter 3 that, for both the Yule and uniform models, asymptotically the mean and variance of the number of cherries is proportional to n. Therefore we can write $\mu_n \sim \alpha n$ and $\sigma_n^2 \sim \beta n$ for appropriate constants. Given this, we have the following lemma.

Lemma 20 *Let $\mu_n \sim \alpha n$, $\sigma_n^2 \sim \beta n$ where α, β are real constants. For the martingale difference*

$$X_n = C_n - C_{n-1} - 1 + \frac{2C_{n-1} + cn + d}{an + b}$$

we have

$$\lim_{n \to \infty} \mathbb{E}[X_n^2] = \frac{1}{a^2}[-4\alpha^2 + 2a\alpha - 4\alpha c + ac - c^2] .$$

Proof. Let $D_n = C_n - C_{n-1}$ be a random variable representing the difference in the number of cherries for a tree on $n-1$ leaves to a tree on n leaves. D_n can only take on the value 0 or 1. Then for the random variable X_n we have

$$X_n = \begin{cases} \frac{2C_{n-1}+cn+d}{an+b}, & D_n = 1, \quad p_n = 1 - \frac{2C_{n-1}+cn+d}{an+b}. \\ -1 + \frac{2C_{n-1}+cn+d}{an+b}, & D_n = 0, \quad 1 - p_n = \frac{2C_{n-1}+cn+d}{an+b}. \end{cases}$$

Since the events $\{0, 1\}$ form a partition of the event space of D_n then

$$\begin{aligned} \mathbb{E}[X_n^2] &= \mathbb{E}[X_n^2|D_n = 0]\mathbb{P}[D_n = 0] + \mathbb{E}[X_n^2|D_n = 1]\mathbb{P}[D_n = 1] \\ &= \mathbb{E}\left[\left(-1 + \frac{2C_{n-1} + cn + d}{an + b}\right)^2 |D_n = 0\right](1 - p_n) + \mathbb{E}\left[\frac{2C_{n-1} + cn + d^2}{an + b}|D_n = 1\right]p_n \\ &= (1 - p_n)\left\{1 - \frac{2}{an + b}\mathbb{E}[2C_{n-1} + cn + d|D_n = 0]\right\} + \\ &\quad \frac{1}{(an + b)^2}\mathbb{E}[(2C_{n-1} + cn + d)^2]\,. \end{aligned} \tag{C.3}$$

The first term of (C.3) may be expanded out to give

$$(1 - p_n) - \frac{4(1 - p_n)}{an + b}\mathbb{E}[C_{n-1}|D_n = 0] - \frac{2(1 - p_n)(cn + d)}{an + b}\,. \tag{C.4}$$

The second term of (C.4) can be rewritten:

$$\begin{aligned} \frac{4(1 - p_n)}{an + b}\mathbb{E}[C_{n-1}|D_n = 0] &= \frac{4(1 - p_n)}{an + b}\sum_k k\mathbb{P}[C_{n-1} = k|D_n = 0] \\ &= \frac{4(1 - p_n)}{an + b}\sum_k k\frac{\mathbb{P}[D_n = 0|C_{n-1} = k]\mathbb{P}[C_{n-1} = k]}{\mathbb{P}[D_n = 0]} \\ &= \frac{4}{an + b}\sum_k k\left(\frac{2k + cn + d}{an + b}\right)\mathbb{P}[C_{n-1} = k] \\ &= \frac{8}{(an + b)^2}\sum_k k^2\mathbb{P}[C_{n-1} = k] + \frac{4(cn + d)}{(an + b)^2}\sum_k k\mathbb{P}[C_{n-1} = k] \\ &= \frac{8}{(an + b)^2}\mathbb{E}[C_{n-1}^2] + \frac{4(cn + d)}{(an + b)^2}\mathbb{E}[C_{n-1}]\,. \end{aligned}$$

Substituting back into (C.4) then we have for the first term of (C.3)

$$(1 - p_n) - \frac{8}{(an + b)^2}\mathbb{E}[C_{n-1}^2] - \frac{4(cn + d)}{(an + b)^2}\mathbb{E}[C_{n-1}] - \frac{2(1 - p_n)(cn + d)}{an + b}\,. \tag{C.5}$$

For large n

$$1 - p_n \sim \frac{2\alpha + c}{a} \qquad p_n \sim 1 - \frac{2\alpha + c}{a}$$

and

$$\frac{\mathbb{E}[C_n]}{n} \sim \alpha \qquad \frac{\mathbb{E}[C_n^2]}{n^2} \sim \alpha^2.$$

Using (C.5) and the above asymptotic results gives for the first term of (C.3), as n becomes large,

$$\begin{aligned}\lim_{n\to\infty}(1-p_n)\left\{1 - \frac{2}{an+b}\mathbb{E}[2C_{n-1} + cn + d | D_n = 0]\right\} \\ &= \frac{2\alpha + c}{a} - \frac{8}{a^2}\alpha^2 - \frac{4c}{a^2}\alpha - \frac{2(2\alpha + c)c}{a^2} \\ &= \frac{1}{a^2}[a(2\alpha + c) - 4\alpha(2\alpha + c) - 2c(2\alpha + c)] \,. \end{aligned} \tag{C.6}$$

The second term of (C.3) may be expanded as

$$\frac{4}{(an+b)^2}\mathbb{E}[C_{n-1}^2] + \frac{4(cn+d)}{(an+b)^2}\mathbb{E}[C_{n-1}] + \frac{(cn+d)^2}{(an+b)^2} \,.$$

This gives, asymptotically,

$$\lim_{n\to\infty}\frac{1}{(an+b)^2}\mathbb{E}[(2C_{n-1} + cn + d)^2] = \frac{1}{a^2}[4\alpha^2 + 4c\alpha + c^2] \,. \tag{C.7}$$

Combining the asymptotic results of (C.6) and (C.7) back into (C.3) gives the final result

$$\begin{aligned}\lim_{n\to\infty}\mathbb{E}[X_n^2] &= \frac{1}{a^2}[a(2\alpha + c) - 4\alpha(2\alpha + c) - 2c(2\alpha + c)] + \frac{1}{a^2}[4\alpha^2 + 4c\alpha + c^2] \\ &= \frac{1}{a^2}[-4\alpha^2 + 2a\alpha - 4\alpha c + ac - c^2] \,. \end{aligned}$$

□

We are now in a position to prove that $\{X_n\}$ satisfy the Lindeberg condition, which given the previous two lemmas is quite straightforward.

Proposition 10 *The random variable*

$$X_n = C_n - C_{n-1} - 1 + \frac{2C_{n-1} + cd + d}{an + b} \,,$$

where a, b, c, d are real constants, satisfies the Lindeberg condition.

Proof. Since X_n can take on more than one value for $n \geq 4$, part (i) of the Lindeberg condition is satisfied. To see why part (ii) holds note that from Lemma 19

$$|X_n| \leq \gamma \qquad \forall n \qquad \text{where} \quad \gamma = 1 + \frac{|c| + |d| + 1}{|a|} .$$

Therefore we can write

$$\begin{aligned} \sum_{x_{jl}:|x_{jl}|>\epsilon\ s_n} x_{jl}^2 p_{jl} &\leq \gamma^2 \mathbb{P}[\ |X_j| > \epsilon s_n] \\ &\leq \frac{\gamma^2 \sigma_j^2}{\epsilon^2 s_n^2} \qquad \text{using Chebyshev's Theorem [27].} \end{aligned}$$

Summing over j, and dividing by $s_n^2 = \sum_{j=1}^n \sigma_j^2$, gives

$$\frac{1}{s_n^2} \sum_{j=1}^{n} \sum_{x_{jl}:|x_{jl}|>\epsilon s_n} x_{jl}^2 p_{jl} \leq \frac{\gamma^2}{\epsilon^2 s_n^2}.$$

From Lemma 20 $\lim_{n\to\infty} \sigma_j^2 = \lim_{n\to\infty} \mathbb{E}[X_n^2]$ is non-zero, therefore $\lim_{n\to\infty} s_n^2 = \infty$. Thus it follows that

$$\lim_{n\to\infty} \frac{1}{s_n^2} \sum_{j=1}^{n} \sum_{x_{jl}:|x_{jl}|>\epsilon s_n} x_{jl}^2 p_{jl} = 0 \qquad \forall \epsilon > 0 .$$

□

Appendix D

List of Symbols

Symbol	*Meaning*	*Page*
c_n	Catalan number	8
C_n	number of cherries	38
T_n	number of triplets	48
$RB(n)$	set of labelled rooted trees on n leaves	7
$[n]$	the set $\{1, 2, \ldots, n\}$	63
$\|S\|$	the number of elements in the set S	4
$\mathcal{S}(n)$	number of unlabelled rooted trees on n leaves	6
$f(T)$	total distance over all $\binom{n}{2}$ pairs of leaves for the tree T	29
D_n	expected value of $f(T)$	29
$d(i, j)$	distance between the leaves i and j	30
$d(i, \rho)$	distance of the leaf i from the root vertex ρ	30
μ_n	mean distance from the root for a tree on n leaves (Yule model)	26
σ_n	variance of the distance from the root for a tree on n leaves (Yule model)	26
ν_n	mean distance from the root for a tree on n leaves (uniform model)	27
d_n	mean distance between two leaves (Yule model)	31
$\mathtt{d_n}$	mean distance between two leaves (uniform model)	34

H_n	number of histories with n leaves	14
$H_n(t)$	number of histories for the tree t	14
$E_{\max}(T)$	set of maximum likelihood edge	54
$e_{\max}(T)$	any edge in $E_{\max}(T)$	54
$\mathcal{L}(T)$	the leaf set of the tree T	5
$(2n-3)!!$	$(2n-3) \times (2n-5) \times \cdots \times 5 \times 3 \times 1$	7
σ	symmetry index of a tree shape	7
$\left[{n \atop k} \right]$	unsigned Stirling number of the first kind	24
$\mathbb{E}[X]$	expected value of random variable X	4
$\mathbb{P}[T]$	probability of the tree T (labelled or unlabelled)	4
$\mathbb{P}_n[T]$	same as $\mathbb{P}[T]$ but emphasising that T has n leaves	4
$\mathcal{N}(\mu, \sigma^2)$	normal distribution with mean μ and variance σ^2	41

Index